



**MARY KAY O'CONNOR  
PROCESS SAFETY CENTER**  
TEXAS A&M ENGINEERING EXPERIMENT STATION

---

20<sup>th</sup> Annual International Symposium  
October 24-26, 2017 • College Station, Texas

---

**Can we verify and intrinsically validate risk assessment results?  
What progress is being made to increase QRA trustworthiness?**

**Hans Pasman\* and William Rogers**

*Mary Kay O'Connor Process Safety Center  
Artie McFerrin Department of Chemical Engineering  
Texas A&M University  
College Station, Texas 77843-3122*

\*Presenter E-mail: [hjpasman@gmail.com](mailto:hjpasman@gmail.com)

**Abstract**

The purpose of a risk assessment is to make a decision whether the risk of a given situation is acceptable, and, if not, how we can reduce it to a tolerable level. For many cases, this can be done in a semi-quantitative fashion. For more complex or problematic cases a quantitative approach is required. Anybody who has been involved in such a study is aware of the difficulties and pitfalls. Despite proven software many choices of parameters must be made and many uncertainties remain. The thoroughness of the study can make quite a difference in the result. Independently, analysts can arrive at results that differ orders of magnitude, especially if uncertainties are not included. Because for important decisions on capital projects there are always proponents and opponents, there is often a tense situation in which conflict is looming.

The paper will first briefly review a standard procedure introduced for safety cases on products that must provide more or less a guarantee that the risk of use is below a certain value. Next will be the various approaches how to deal with uncertainties in a quantitative risk assessment and the follow-on decision process. Over the last few years several new developments have been made to achieve, to a certain extent, a hold on so-called deep uncertainty. Expert elicitation and its limitations is another aspect. The paper will be concluded with some practical recommendations.

**Keywords:** Risk assessment, uncertainty, trustworthiness, decision making

**1. Introduction**

Process safety involving hazardous materials can be evaluated only by determining remaining risk. As we all have the task to do the utmost to prevent the nasty surprise of a major accident, we must perform risk assessments one way or another. Now, in most cases we can rely on situations we know, and we can estimate risk magnitudes at least in ranges. In case there are quite a few risk

sources around, with the aid of a risk matrix we can perform a semi-quantitative assessment on a comparative or even an order of magnitude basis, in which we can visually overlook and categorize the risk situation. However, difficulty arises when the potential consequences can be catastrophic, costly risk reduction becomes an issue, and the case must be considered on a quantitative basis, a quantitative risk assessment or QRA, using distributions instead of only ranges for consequence and frequency or probability of occurrence. Of course, in such cases the probability of occurrence will generally be very low. This constitutes the typical rare event prediction problem. Such cases have given rise to endless debates and are the topic of many scientific papers, because experimental validation of high impact, low probability risk is virtually impossible. The final result of an assessment depends strongly on assumptions and the data used. Depending on what is judged reasonable as starting material, different assessors with different background knowledge can come to end results that in a quantitative sense can differ by orders of magnitude especially if uncertainty is not properly assessed. Regretfully, the practice to discuss uncertainties in studies and to present ranges of inaccuracy is essential but still uncommon.

Due to risk assessment results afflicted with uncertainty, parties with opposing interests have been fighting in many forums to get a project accepted or not, and often when it gets into the public arena such project becomes a political bone of contention. As Aven and Zio [1] note: “The disguised subjectivity of risk assessments is potentially dangerous and open to abuse if it is not recognized.” And also: “Precise numbers are used as a facade to cover up what are often political decisions”. The work procedure for a risk assessment sounds easy but the execution is usually rather problematic and one can easily lose trust in the result. Predicting risk in a given situation with many hazards present and with the many possibilities in which a scenario can conditionally develop, is in a physical sense, regarding the range of possible consequences, not that simple but in probabilities almost impossible. Yet, we are depending on assessments to support optimum decision making and weighing of effort against cost. And also, it is not just a specific problem of process safety. Risk management in general, whether it is on project planning and execution, financial or business, all face the same problem. We want to look into the future, but predictive tools are fallible and the prediction results are helpful but uncertain.

In fact, the problem has two sides: one side is the thoroughness of QRA treatment and assessment of uncertainty, and the other is what can we do to validate a result and how much can the result of an analyst team, given their capabilities, be trusted.

In this paper, we shall look at these aspects, and we shall start in Section 2 with an extreme, but for certain manufacturers serious case in which a team or company is asked to give more or less a guarantee on the frequency of a rare event prediction result. This is related to a product of which the user can be killed in the event. In Section 3 we shall briefly review some ways to deal with uncertainty in process risk. In Section 4 before reaching a conclusion, we shall indicate in what directions we can improve and validate, and we will discuss what can be done to obtain greater confidence in view of decision making under uncertainty and policy developing.

## **2. The Safety Case trustworthiness**

First the designation “Safety case”: In the U.S. it is used often as an equivalent to ‘Risk assessment’, but that is not fully right. A ‘safety case’ is quite strict following the definition in the U.K. Defence Standard (DS 00-56) [2] as “a structured argument, supported by a body of evidence that

provides a compelling, comprehensible, and valid case that a system is acceptably safe for a given application in a given operating environment”. Although the statement is entirely qualitative, the target is in most cases quantitative, and the choice of words does not leave any doubt that it is a serious and stringent requirement to satisfy. As a term, ‘safety case’ became known to indicate a risk assessment and a report that should convince the competent authorities in the countries around the North Sea that the offshore operation is safe. In the U.K., according to HSE, The Health and Safety Executive, the Offshore Installations (Safety Case) Regulations 2005 (SCR05) [3] aim to reduce the risks from major accident hazards to the health and safety of the workforce employed on offshore installations and in connected activities. The regulations implement the main recommendations of Lord Cullen's Report of the Public Inquiry into the 1988 Piper Alpha Disaster. However, the term also became more generally applied, even according to the meaning the defense standard intends. This is to give products that in case of failure could result in death of the user or in some type of disaster, confidence that use of the product is safe, e.g., regarding aircraft, pacemaker, safety-critical instrumentation and software, et cetera. The above mentioned standard DS 00-56 requires evidence based assuring arguments and the confidence in a claim that the system is acceptably safe. So, if the claim is that the pacemaker will have a maximum failure rate of  $10^{-5}/\text{yr}$ , the manufacturing company must provide persuading evidence for this claim.

Sujan et al. [4] in connection to the question whether health care should adopt the concept of the safety case, provide more extensive background and development of the safety case and of the various industry branches in the U.K., where it has become practice. Meanwhile, the concept has become in use also at the U.S. Department of Defense for certain projects and at NASA.

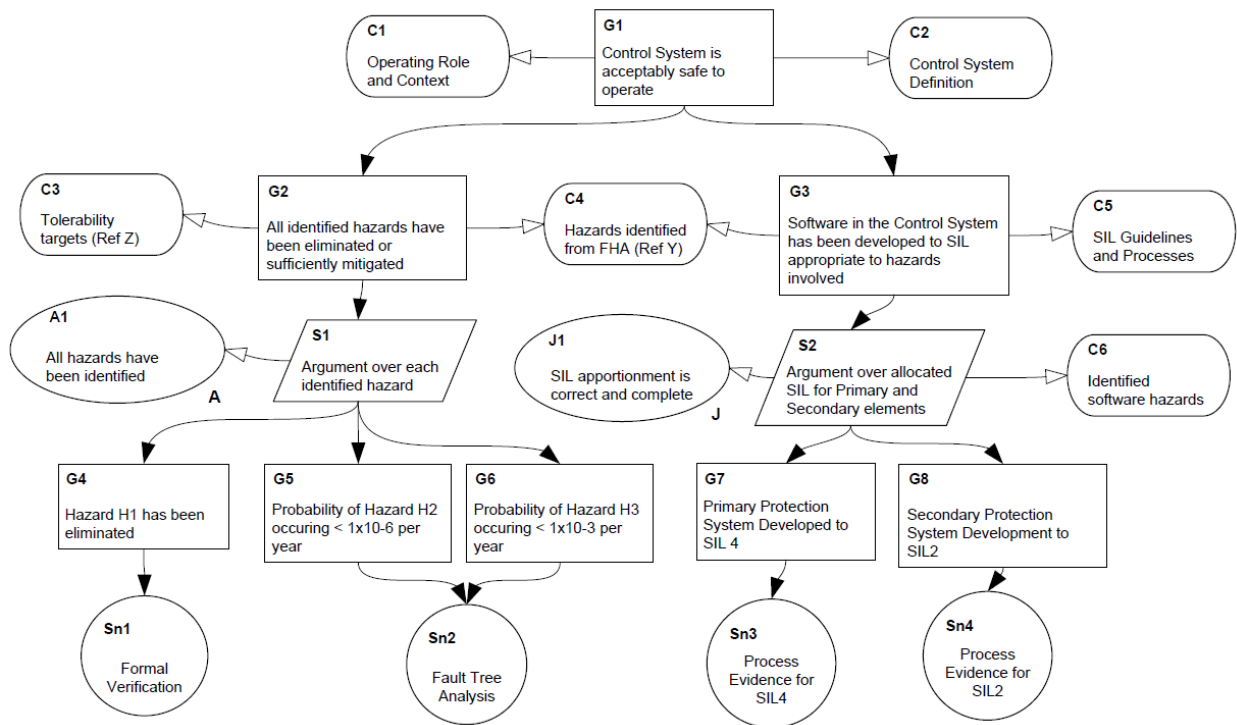


Figure 1. Example of Goal Structuring Notation (GSN Community [5]) for a safety case of a SIS control system. Rectangular means Goal, G; a circle is a Solution, Sn; a rhomb is a Strategy,

S, linking lower goals to a higher one, and connected to an oval being either an oval Assumption, A, or Justification, J; finally, the tank like balloons are Comments, C.

The assurance process is goal based: It sets requirements but does not state how to fulfill these requirements. The University of York, U.K., with among others, Tim Kelly, is an important player in developing the assurance concept and in organizing a community of interested industry members. In 2011 the GSN Community [5] proposed the Goal Structuring Notation (GSN) Standard as a means to underpin assurance arguments. The Standard states: “GSN is a graphical argumentation notation that can be used to document explicitly the individual elements of any argument (claims, evidence and contextual information)”. The definition of an assurance is given as: “A reasoned and compelling argument, supported by a body of evidence, that a system, service or organisation will operate as intended for a defined application in a defined environment”. Further an argument is defined as: “a connected series of claims intended to establish an overall claim”. The argued claim can be qualitative or quantitative and is deemed true or false. An example of a GSN graphical structure related to a Safety Instrumented System is given in Figure 1. The standardized graphical structure serves the communication among the different stakeholders. There is much more to GSN than described here, but Figure 1 gives an impression. Quite a few companies have adopted GSN.

A distinction is made between a *safety* argument and a *confidence* argument, which is made to back the safety argument. The former spells out the asserted arguments for reducing the risk to the residual risk level, while the latter presents the reasons why one can have confidence in that result. For the safety argument, the ALARP (As Low As Reasonably Practicable) within the acceptable range criterion requirements can be used. In a report by Nair et al. [6] an approach is described of what is called *evidential reasoning* to assess the confidence one can achieve in the safety argument. This latter concept is why we mention this here, because that is what should produce or increase credibility and trustworthiness. The principle is building a hierarchy of attributes, which depending on importance can be weighted, while the base level of attributes is assessed by an expert on, e.g., a 5-point Likert-scale. By introducing subjective probability as a mathematically less strict belief function, an expert can express his/her uncertainty by attributing a belief value of say 25% to the highest scale “excellent” and 75% to “good”. The beliefs need not sum to 100%, and if the assessor has no knowledge about an attribute, zero is submitted. The belief function is the core of the Dempster-Shafer theory of evidence to be further explained in Section 3. A number of rules have been developed to assimilate the attributes. Nair et al. [6] developed a software tool (EviCA = Evidence Confidence Assessor) for guiding a user and aggregating the various sub-claims in a coherent final assertion.

The structure of attributes built for the confidence argument makes a main distinction between *trustworthiness* and *appropriateness*. *Trustworthiness* is subdivided into Personnel, Process/Techniques, Tool Integrity, Content Compliance, Evidence Past, and a User-Defined Trustworthiness Factor. *Personnel* is further subdivided into Past Knowledge, Competency, Independence, Domain Experience, and User-Defined Personnel Factors; *Processes/Techniques* into Past Use, Definition, Peer Review, and User-Defined Processes/Techniques Factor; *Tool Integrity* into Bound Qualification, Standard Qualification, and User-Defined Tool Factor; *Content Compliance* into Scope, Expected structure, and User-Defined Content Compliance Factor. Then, the other top factor *Appropriateness* is sub-divided into Scope, Expected Structure, and User-Defined Content Compliance Factor. The meaning of these factors is explained in the report of

Nair et al. [6], and shown in GSN graphs. To mention all that here is to give an impression of the thoroughness of the assessment. The report describes the use of a software tool and some examples.

Besides the mentioned method of Evidential Reasoning there are quite a few other methods in use for the same purpose. In a very interesting and elaborate paper by Graydon and Holloway [7], an overview of all the techniques in use is presented and a test performed to investigate whether a technique can yield a positive assessment of a case in which it would not have been justified. Twelve different teams practicing safety case assurance were identified. From these, five applied Bayesian networks, six the related methods of Dempster–Shafer Theory, Jøsang’s Opinion Triangle, or Evidential Reasoning, and one based on simple weighting, proposed by Yamamoto.

In case of the Bayesian Belief networks (BBN), various attributes are represented by network nodes and the assimilation by edge connections between nodes producing a cause-effect network structure. The parent base nodes are fed the attribute subjective belief probabilities, upon which the top leaf node produces the confidence degree. However, all five teams applying BBN structure of the network calculated the node probability value differently.

Dempster-Shafer theory not only measures the strength of a belief probability but also the plausibility. Jøsang’s Opinion Triangle discerns three dimensions: belief, disbelief, and uncertainty, and Evidential Reasoning has been briefly explained above and functions according to the Dempster-Shafer theory. Analysts following Yamamoto rate GSN attributes on a five-point Likert scale from strongly unsatisfied to strongly satisfied and assess the overall confidence in the claim by means of weighted averages.

Graydon and Holloway [7] analyzed all 12 methods in detail and applied on each a counterexample inspired by the example used by the authors describing the assurance technique. In all cases a counterexample could be identified in which the technique’s output was implausible, and hence, the technique proved to be not fully trustworthy. Detailed results can be found in a NASA report (Graydon and Holloway [8]). This result does not mean that the techniques used are wrong, but as Graydon and Holloway concluded the techniques are imperfect. It only shows how difficult it is to prove *decisively* an assessment result of the occurrence probability of a rare event to be below a certain low value. Instead, we should focus on estimating the *credibility* of an occurrence probability less than a selected acceptably low value.

### **3. Treatment of process and plant risk uncertainty**

Over the years much has been learned on risk assessment and how to treat uncertainty, but relatively little of the new learning has become common practice. This year a special issue of the journal of Safety Science on the topic of Risk Analysis Validation and Trust in Risk Management will appear. The present authors contributed an article entitled: “Risk assessment: What is it worth? Shall we just do away with it, or can it do a better job?” (Pasman, Rogers, Mannan [9]). Of course, the answer to the rhetoric question is “yes”, but it takes an effort. We shall briefly summarize the findings.

First, a positive note about *consequence analysis*, which has been much improved in methods and performed within widened applications. Yet, there is still uncertainty. MKOPSC accomplished much learning through fundamental work in outflow and boiling of cryogenics, such as LNG, but

prediction of evaporation rate in a given case is still not a clear-cut task. A few other examples: calculation of dispersion of cloud by means of different 3-D computational fluid dynamics models gave for a number of aspects an improvement over previous integral models, but for good confidence the spread in results for a given case obtained using different codes is still too large. The low wind speed condition, in which most damage effect is possible, requires more attention. Although more and more is known about the BLEVE phenomenon, we would like to know more about the details of time and effect distribution of this physical explosion behavior, also in view of decision making during emergency response action. There has been much progress made in knowledge about vapor cloud explosion, and now it is understood that given certain conditions a large flammable cloud can detonate; further development of models is required. Finally, the number of probit relations on various effect phenomena describing probability of extent of damage is still growing and expanding also with respect to domino effects. Unfortunately, including uncertainty ranges in results is also in consequence analysis still uncommon.

However, the most serious problems in risk assessment are posed by *hazard identification*, *scenario definition*, and *failure rate data*. On invitation and in cooperation with Prof. Cameron and colleagues in Australia a review has been written about problems and perspectives regarding HAZID methods (Cameron et al. [10]). Due to the many possibilities something can go wrong, scenario completeness for a given plant is not easy to achieve and certainly not the assurance the scenarios identified are complete. But, adoption of a system approach and use of computerization are the way to achieve a more comprehensive hazard identification and possible scenarios for a socio-technical system representing an operational plant and its organization.

Failure rate uncertainty is an old problem. Although some data bases are available, applicable data are scarce and conditions of failure badly defined, so that the well-developed framework of equations describing availability cannot be used to its full potential. Lack of willingness out of fear to public exposure and legal consequence hinders cooperation with respect to failure rate data and impedes progress. Risk assessment pioneers and proposers of the triplets of risk analysis: scenario - failure frequency – consequence ( $s, \lambda, x$ ), Kaplan and Garrick [11], recognized already in 1981 the uncertainty problem, and launched the *probability density of frequency* concept. If data are collected on the failure frequency of an equipment one can build a distribution  $p(\lambda)$  and find an average, median, and variance. Data are scarce, especially when it concerns a particular brand and type of equipment. In fact, one can apply this reasoning also on scenario and consequence. Kaplan and Garrick therefore proposed the *Bayes theorem* stating that prior distribution multiplied by a likelihood distribution of new evidence while normalizing the product to a maximum probability of unity, yields a posterior distribution. Bayes theorem allows inclusion of data from similar equipment and plants, hence not strictly data from the same population. The uncertainty the Bayes model aggregates will be expressed again in the variance and shape of the posterior. Now, more than three decades later we are teaching and applying the principle, but its use in actual practice of process safety is still very reluctant, yet it is needed now more than ever. An example how it works, is given in the Appendix.

For the rare event frequency, Kaplan and Garrick [11], already in 1981 much advanced in thinking, mentioned the *multiple-stage* use of the Bayes theorem, formally called *Hierarchical Bayesian Analysis* (HBA). Suppose following a certain initiating event (IE), one can notice an alert, or barrier functioning, or a failure that could have been developed to a serious accident, but thanks to safeguarding or another mechanism the accident didn't materialize. Suppose further, that

as indicators one can record over a certain period of, e.g., 20-30 years the number of these precursors and their corresponding end states. If both indicators and end-states are specific for the IE, these data can be used for rare event frequency prediction. The data can be even from the sector as a whole instead of own plant only. In such case an event tree representing the possible physics of the phenomena should be developed. This tree branching out following a particular IE links the sequence of successive at the branch points observable events in the tree (e.g., vibration, corrective maintenance, smoke, alarm), to counted precursor end states of damage (e.g., no damage, trip, small leak, large leak and emergency shutdown, fire, disastrous explosion), which also can be expressed in monetary loss values. The observed number of end state occurrences of each type  $i$ ,  $n_i$  in a given time,  $t$  with an average occurrence rate estimator,  $n_i/t = \lambda_i$  can be modeled, e.g., as a Poisson distribution, which is an exponential function:  $exp(-\lambda_i t)\{(\lambda_i t)^{n_i}/n_i!\}$ . Each type  $i$  will have a different but related sector source. The source-to-source variability for  $\lambda_i$  is modeled in the first stage of a two-stage HBN with an aggregated Poisson likelihood for  $\lambda_i$  and a prior gamma distribution,  $\pi(\lambda_i|E) = \iint \pi_1(\lambda_i|E, \alpha, \beta)\pi_3(\alpha, \beta|E)d\alpha d\beta$  conjugate to Poisson, i.e. mathematically compatible and producing a posterior gamma distribution. In this equation  $\alpha$  and  $\beta$  for the second stage of the HBN are so-called hyper-parameters assumed to be independent. Given as a worst case of no prior information on  $\alpha$  and  $\beta$  at all, a uniform distribution representing an uninformative hyperprior, but preferably when any information is available with an informative hyperprior, for example an expert estimate, the theorem will produce a posterior in  $\alpha$  and  $\beta$ .

This posterior will be based too on the evidence of  $n$  and  $t$ , which instead of fixed values are now functions of  $\alpha$  and  $\beta$ . In the second step for a particular precursor type a predictive density for  $\lambda_i$  is derived as  $\pi(\lambda_i|E) = \iint \pi_1(\lambda_i|E, \alpha, \beta)\pi_3(\alpha, \beta|E)d\alpha d\beta$ , in which  $\pi$  is a gamma probability distribution, and  $E$  is evidence. The solutions to the equations are obtained by WinBUGS software based on the Markov-Chain Monte-Carlo technique. The result takes account of the variability among sources, and by taking data of a specific plant in the sector via a conventional Bayesian update the plant specific  $\lambda_{i,sp}$  are found. Also, the aggregate of the posterior means, the  $\lambda_{i,sp}$ , for each tree sequence yields the IE frequency, while the ratios resolve the branch probabilities of the event tree. Given later new evidence, the Bayes theorem can further update the rate. If, as usual the case, the same end-state, e.g., a trip, can be reached via different branches of the tree a different route has to be followed. On demand the top events are binomial success or failure. Their numbers can be derived from the end state ones, while IE is the sum of all numbers. Then, the estimator is  $\hat{p}_j = n_j/n_{IE}$ . For binomial distributed variables, the hyper-parameters are assumed to be beta-distributed, where beta is conjugate to binomial. The further calculation proceeds similarly to the above. The event tree can also be modeled as a Bayesian network. The principle has been demonstrated the last few years in a few papers (Yang et al. [12] and [13]; Khakzad et al. [14]), which line out further possibilities and applications.

Uncertainties with respect to possible scenarios, and consequence estimates also deserve much attention. If a bowtie is constructed, and even better a Bayesian network using a bowtie as a starting point, and input parameter uncertainties are specified, these uncertainties will be propagated throughout the network, and the effect on the spread of the risk result can be determined. The uncertainties can be specified either, e.g., as variance of truncated normal distribution of an input parameter or as a uniform or flat distribution, indicating only an expected interval. In fact, uncertainty analysis can be conducted also as sensitivity analysis by finding out which variables, parameters, and uncertainties have the largest effect on the final result.

A considerable number of major accidents occurred according to scenarios that had been excluded before as improbable. Usually it is cost of measures that drives the decision and the large uncertainty margin of the probability of occurrence. Yet, the larger the potential consequence, the lower a failure event probability or frequency of occurrence should be. Nuclear power plants are the best-known example. The Fukushima board was warned and knew that in case of high tsunami, the reactors would be threatened. So, how can we make good decisions in risk management, or rather how to make optimum decisions under uncertainty? Much has been written about the topic and methods have been proposed in relation to economics and investment decision. More specifically relevant to industrial risk there are rigorous methods applying probabilities, others based on approximate reasoning, and finally using only expert knowledge. In all, background knowledge and belief play a role. Most rigorous is the Bayesian approach just mentioned, which can also make use of subjective probability as input. The latter is the kind of probability one would quote when asked how likely it will rain tomorrow. The Bayes model is not biased by preferences, and it analyzes evidence from events but not the events themselves. Bayes is the most coherent and effective method to analyze information about events, propagate the uncertainties, and support evidence based reasoning and decisions with estimated uncertainties. If cost factors are introduced in a Bayesian network, the annual average expected loss can be calculated and be part of business decision-making. (Pasman and Rogers [15]).

Less rigorous but still close to Bayesian belief functionality and known for managing uncertainty is the *Dempster-Shafer* method. In fact, the Evidential Reasoning described in Section 2 is an example of it (Zeng et al. [16]; Nair et al. [6]). The method is about degrees of belief and subjective probability and as a method of approximate reasoning that is suited in cases of conflicting or confusing evidence when the rigorous Bayesian method does not offer a solution. Shafer's 1976 contribution [17] is to distinguish your subjective probability estimate of how reliable an expert is in general, and his/her degree of belief whether a certain event or issue occurs. The degree of belief the event/issue does *not* take place, can be a value smaller than the difference of occurrence probability, and can even be set to be zero, meaning only that the expert did not have a reason why the event or issue should not occur. The gap between 1 and the sum of both probabilities for occur & not occur, is called *ignorance*, which discerns it from probability theory where it is always 0. Occurrence *plausibility* is 1 minus belief of non-occurrence. Degree of belief in evidence is analogous to mass,  $m$  assigned to probability. Now, there could be related but independent evidence in the environment, e.g., amplifying what the expert states, or there is another independent expert statement, who even in part may contradict that of the first one. Dempster's 1968 combination rule [18] enables fusing belief functions, e.g.,  $m_1$  and  $m_2$ , and is written as follows:

$$m_1 \oplus m_2(Z) = \left(\frac{1}{K}\right) \sum_{X \cap Y = Z} m_1(X)m_2(Y); \quad K = 1 - \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y)$$

$X \cap Y = Z$  includes all those elements that  $X$  and  $Y$  have in common, while  $\oplus$  means summing the products of intersecting terms at the right side of the equation.  $K$  is a normalization factor and has the value 1 for the example that follows below. The new mass represents consensus as it consists of the agreement parts of the inputs and not the disagreement ones. A simple example is the following. A question is whether in a given plant situation a critical event can occur or not. An expert, thought to be highly reliable, estimates it as possible with 75% chance. A second expert, also renowned, considering modernization estimates the possibility only 30%. The combination rule for Expert 1 OR Expert 2 assuming independence of experts produces  $1 - (1 - 0.75)(1 - 0.30) = 0.83$  degree of support that the event is possible (unity minus the product of the non-assigned

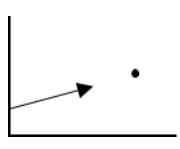
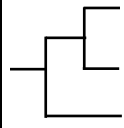
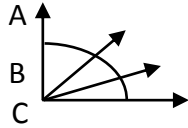
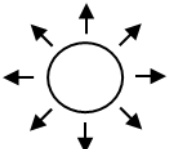


beliefs of the two experts, that is what they have in common, yields new belief). It can be performed with three elements or more, e.g., there had been an accident 20 year ago or not, and did the expert know that, but it quickly becomes complicated to formulate the problem correctly. In any case, the Dempster-Shafer method can also be used to generate imprecise data, for example, for a fault tree analysis (Curcurù et al. [19]).

Several other methods have been proposed always with inputs from experts to at least obtain an expectation. Zadeh's *Fuzzy logic*, on which there exists a wealth of literature, with its membership functions has been used extensively, but it has the fundamental shortcoming that in the last step, the defuzzification to obtain a so-called crisp value, all uncertainty information is deleted and an unjustified and illusory feel of certainty can arise. The Analytic Hierarchy Process, also much in use, is based on preferences and is not suitable for engineering decision making. Hazelrigg [20] explains the reasons.

*Expert elicitation* is not a simple matter, and there has been quite some discussion. Bolger and Rowe [21] discussed with respect to the classical method, the best way "to calibrate" the experts, the problems of unreliability, and the aggregation of expert opinion by weighting, called mathematical aggregation, and alternatively behavioral aggregation. In the latter, experts in a group with a facilitator attempt in rounds of deliberation and voting to move toward and come to a consensus. Cooke [22] with a long-time experience in expert elicitation and Bayesian statistical result treatment (see Paskan, Rogers [23]), wrote a critique on the Bolger and Rowe paper. The focus of the critique is on the calibration and type of weighting: equal weighting or differential, also called performance weighting, of which Cooke is a proponent. (In their response, Bolger and Rowe [24] stress that it had been their intention to discuss the advantage of behavioral aggregation over equal weighting and not to discuss performance weighting!)

Petri net is attractive for random, safety critical timing problems, but it is rather effort intensive, because the software requires additional detailed programming.

		Level 1	Level 2	Level 3	Level 4	
		Deep Uncertainty				
<b>Determinism</b>	<b>Context</b>	A clear enough future 	Alternate futures (with probabilities) 	A multiplicity of plausible futures 	Unknown future 	<b>Total Ignorance</b>
	<b>System Model</b>	A single system model	A single system model with a probabilistic parameterization	Several system models, with different structures	Unknown system model; know we don't know	
	<b>System outcomes</b>	A point estimate and confidence interval for each outcome	Several sets of point estimates and confidence intervals for the outcomes, with a probability attached to each set	A known range of outcomes	Unknown outcomes; know we don't know	

<b>Weights on outcomes</b>	A single estimate of the weights	Several sets of weights, with a probability attached to each set	A known range of weights	Unknown weights; know we don't know	
----------------------------	----------------------------------	--	--------------------------	-------------------------------------	--

Figure 2. Taxonomy of uncertainties according to Cox [25] after an original by Walker et al. [26]

When knowledge is lacking (epistemic) *deep uncertainty* is most problematic. Cox [25] provides guidance with the scheme given in Figure 2. First, a taxonomy of degrees of uncertainty and lack of knowledge makes clearer where deep uncertainty starts. If in the setting of a Level 1 risk management, decisions must be made on alternative risk reducing measures, what Cox calls acts, the usual approach is to weigh consequences in (dis-)utility terms against acts, which in the end will mean investment. This is called the Subjective Expected Utility, or SEU decision theory. The model of the system is defined and validated. In the decision, probability, e.g., due to parameter value uncertainty is taken into account besides consequences, act options, and the utility. This approach will still work in case of Level 2 uncertainty.

Table 1. Ten methods for decision making under uncertainty after Cox [25]. From top to bottom methods become more rigorous due to increasing uncertainty in depth.

Method	Model generation	Optimization/Adaptation	Combination
Expected utility/SEU theory	One model specified	Maximize expected utility (over all acts in the choice set, $A$ )	None
Multiple priors, models, or scenarios; robust control, robust decisions	Identify multiple priors (or models, or scenarios, etc.) e.g., all models close to a reference model (based on relative entropy)	Maximize the return from the worst-case model in the uncertainty set	Penalize alternative models based on their dissimilarity to a reference model
Robust optimization	Use decisionmaker's risk attitude, represented by a coherent risk measure, to define the uncertainty set	Optimize objective function while satisfying constraints, for all members of uncertainty set	None
Average models	Use multiple predictive (e.g., forecasting) models	None	Simple average or weighted majority
Resampling	Create many random subsets of original data and fit a model to each	Fit models using standard (e.g., least squares, maximum likelihood) statistical criteria	Create empirical distribution of estimates
Adaptive boosting (Adaboost)	Iteratively update training data set and fit new model	Reweight past models based on predictive accuracy	Use weights to combine models
Bayesian model averaging (BMA)	Include all models that are consistent with data based on likelihood	Condition model probabilities on data	Weight models by their estimated probabilities
Low-regret online decisions	Set of experts, models, scenarios, etc. is given, $\{M_1, M_2, \dots, M_n\}$	Reduce weights of models that make mistakes	Weighted majority or selection probability
Reinforcement learning (RL) for MDPPs: UCRL2	Uncertainty set consists of confidence region around empirical values	Approximately solve Bellman equations <sup>1</sup> for most optimistic model in uncertainty set to determine next policy	Update from episode to episode based on new data

Model-free reinforcement learning (RL) for MDPs: SARA	No model used (model-free learning)	Approximately solve Bellman equations <sup>1)</sup> for unknown model	Update value estimates & policies based on new data
---	-------------------------------------	---	---

<sup>1)</sup> Bellman dynamic programming equation is optimization by dynamic programming for solving a complex problem by breaking it down into a collection of simpler subproblems and storing solutions.

However, if model uncertainties arise, Cox identified obstacles that will hamper applying SEU. Model and model parameter uncertainty will cause uncertainty about what acts are best, because consequences, scenario probabilities, utility parameter values, and preferences are uncertain. For these cases of deep uncertainty Cox discusses ten tools, which may lead to better understanding and decision making even when the risk model is uncertain. The ten tools of so-called robust risk analysis are summarized in a table, here reproduced as Table 1, and from the top towards the bottom suited for Level 1 to Level 4 uncertainty. Robust risk analysis means that based on available knowledge and data a number of models/scenarios is generated, which shall be improved and optimized as far as possible and then in some way combined. Such a combination can be performed according to a combination rule: weighting or voting. In Cox [25] each method is discussed in more detail and a few examples are presented of general nature risks, such as climate change.

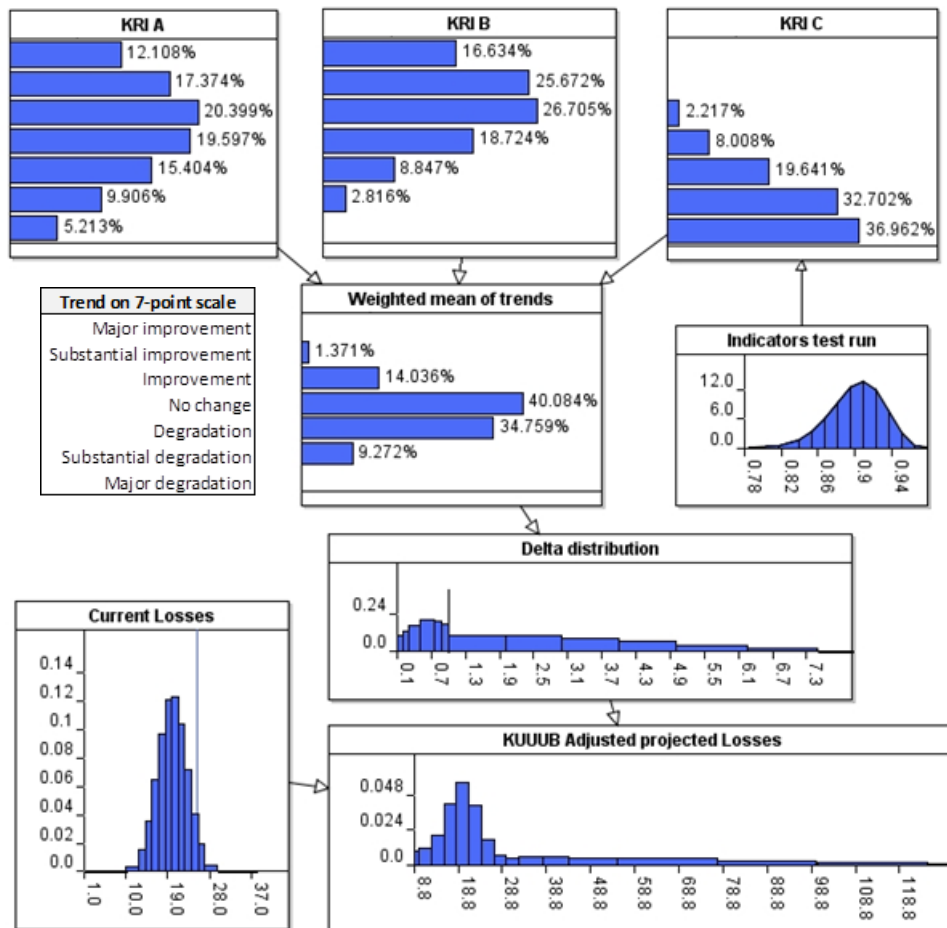


Figure 3. KUUUB factor example adapted from Fenton and Neil [27].

Not the same as the Bayesian model mentioned in Table 1 but related, is the Bayesian network method based on estimates of ‘knowns’ and more or less ‘unknowns’, called the KUUUB factor described by Fenton and Neil [27], see Figure 3. KUUUB is an acronym for Known-Unknown, Unknown-Unknown, and Bias. An adapted example is a company, in which the financial losses due to undesired risk events of an activity A for the current year are not precisely known yet, but are estimated as a truncated normal distribution with mean 20 and variance 10. It is asked to make an estimate for next year’s budget in view of two new activities, B and C. Hence, there are three risk scenarios, called Key Risk Indicators (KRIs). These are expressed as probability of degrees of improvement or of degradation trend on a 7-point scale. KRI A is the known product line with existing risk (weight 2.5); KRI B is an identical line manufacturing a new product with different hazardous properties (weight 1.8) and KRI C is a new high hazard plant (weight 1.0). The distribution of KRI C is conditional on the precursor indicators of a previous test run. By multiplying the weighted mean distribution estimate,  $E$ , of the three KRI trend scenarios with a Delta distribution,  $\Delta$ , a KUUUB adjusted estimate is obtained:  $K_B = \Delta \cdot E$ . Delta parameter-values expressing degrees of uncertainty are conditional on each trend qualification according to Table 2. The parameters are selected based on experienced *expert judgment*. The crux is that Delta is partitioned per trend qualification from major improvement to major degradation. Each partition is modeled as a truncated normal distribution (TND) with mean, variance, upper and lower bound. If there is no change the  $\Delta$ -value collapses to unity (or zero for variance). The compound Delta emerges as a distribution as shown in Figure 3.

Table 2. Delta truncated normal distribution (TND) parameter values conditional on trends

<b>Trends \ TND parameters:</b>	<b>Mean</b>	<b>Variance</b>	<b>Lower</b>	<b>Upper</b>
Major improvement	0.1	0.2	0.1	1
Substantial improvement	0.5	0.1	0.1	1
Improvement	0.7	0.1	0.1	1
No change	1	0	1	1
Degradation	2	4	1	10
Substantial degradation	5	4	1	10
Major degradation	8	2	1	10

The KUUUB adjusted projected loss distribution is less steep than the current losses, because due to uncertain degradation significant probability density mass is drawn into a long uncertainty tail increasing probability of unexpected upsets and greater losses! The indicator node represents aggregated indicators of interdependent socio-technical characteristics, such as emergent behavior and organization resilience, which correspond to changes in thickness of the uncertainty tail. When organization resilience is increased, such as by strengthening the safety culture, the uncertainty tail becomes progressively thinner and low frequency-high consequence events become less likely. The KUUUB model uncertainty tail can therefore be considered a high-level indicator for the viability of the socio-technical organization.

#### 4. Practical measures to validate risk assessment, and to increase confidence

A first possible practical measure is validation of models. Initiated in Europe, 2015, the project SAPHEDRA aims to evaluate and validate consequence models to be used in risk

assessment. For an overview of projects and partners, see SAPHEDRA [28]. Meanwhile, the number of work packages of example applications has been increased to seven. Model evaluation protocols for consequence models should follow an established structure of pre-evaluation tasks, scientific assessment, user-oriented assessment, verification (checking computer implementation of the model), validation (comparison with experimental results), sensitivity and uncertainty analysis, and post evaluation tasks. A first result regarding gas/vapor dispersion models was published in SAPHEDRA HSE [29]. The investigation has reviewed various existing protocols and written recommendations for the structure of a future protocol.

A second effort would be to follow-up the maturity model developed in 2014 by Rae et al. [30]. These authors consider Quantitative Risk Assessment (QRA) to be an engineering method, which means QRA must be evaluated by scientific methods that include being judged on its usefulness and cost effectiveness. For being true the latter demands that an expensive, elaborate, time-consuming QRA shall lead to safer systems with fewer and less costly upsets. This can be turned around: if QRA could be trusted, the safety of a system can be judged. The aim of developing the maturity model is to have guidance to discern an acceptably good QRA from a bad QRA. To make the distinction Rae et al. [30] analyzed a large variety of possible flaws in a QRA. They classified QRA flaws in four levels, followed by a further extensive breakdown of numbered flaws in tables and discussed in special sections. The content of the tables will be briefly summarized here:

Level 1 - *Unrepeatable*: the record/report/description/documentation is incomplete, and it is not possible to reconstruct the assessment. This appears as failure to describe adequately source material, uncertainties, scope, and objectives/evaluation criteria, methods applied, while conclusions and recommendations are ambiguous, incomplete, and when required, are not quantified.

Level 2 - *Invalid*: the effects of flaws in the assessment are larger than the underlying uncertainties in the events investigated, hence “the noise is larger than the signal”. This flaw is with respect to the source data and assumptions. Data can be flawed or available data are not used. In identifying scenarios, external effects, human, organizational, and software failures are not considered. This holds too for non-normal operations, while causal pathways analyzed are incomplete. Further, there is mismatch between assessment and reality, e.g., by wrong assumptions, incorrect application of models and data, and lack of internal consistency of assumptions and models. Then, in the evaluation of results acceptance criteria, such as ALARP, are not correctly applied, costs are not correctly calculated, and alternatives are not considered. Conclusions can be misleading and limitations and uncertainties are not reported. Overall, the analysis is not well conducted: stakeholders and experts may not have been consulted, results may be unrealistic or even worked to an acceptable result, there was no peer review, and the analysis was performed on the wrong issue.

Level 3 – *Valid but inaccurate*: Rigor in selecting data is insufficient: available generic and specific data of better quality have not been used, selection and rejection rules/considerations have not been described; historical data were used without scrutiny of applicability, human and software errors estimates were wrongly determined, expert elicitation was not according strict procedure. Next, are various kinds of errors in applying the models, the use of probabilistics, and performing the calculations. Uncertainties have been given insufficient attention, uncertainties have been inadequately characterized or were not well expressed. Conclusions and

recommendations are inadequately discussed in view of context, assumptions, model, and data limitations.

Level 4 – *Accurate but challengeable*: A risk assessment to be tested is in the first place with respect to disputable data sources, their selection, relative weights, extrapolation, and interpretation. To be insufficient in scientific knowledge on a large variety of aspects, such as assumptions, failure/damage mechanisms, uncertainties of observations, and lacking in scope of existing studies compared to what is needed.

Exceeding level 4 one would achieve an ideal assessment. The maturity model was tested by the authors for completeness by searching literature on errors in risk assessments, further by checking nine peer reviews from different sources involving very different risks, and by their own experience. Realism was tested by making an inventory of the flaws identified in the peer reviews demonstrating that each flaw occurs. Appropriateness was verified by investigating that flaws at level 1 will effectuate that checking flaws at level 2 does not make sense anymore, etc. At the higher levels, an experienced analyst can still make use of the results in a limited way. The maturity models can be used also as a roadmap for research on improvement.

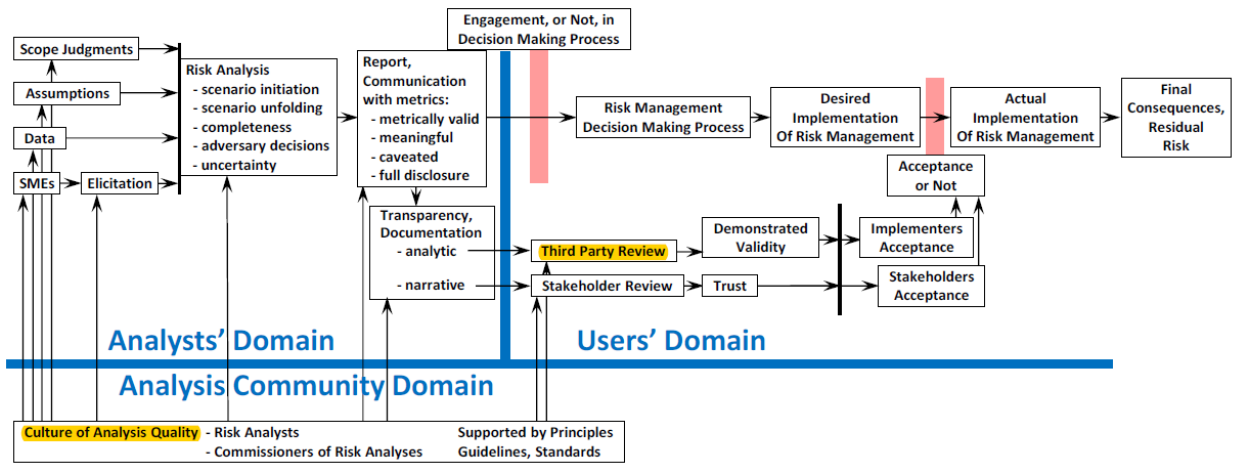


Figure 4. Lathrop and Ezell [31] representation of activities in the three domains of risk analysts, their professional community and the risk assessment users. The highlighting of the boxes of Culture of Analyst Quality and Third-Party Review is by the present authors, because these elements are considered as most essential in the whole process.

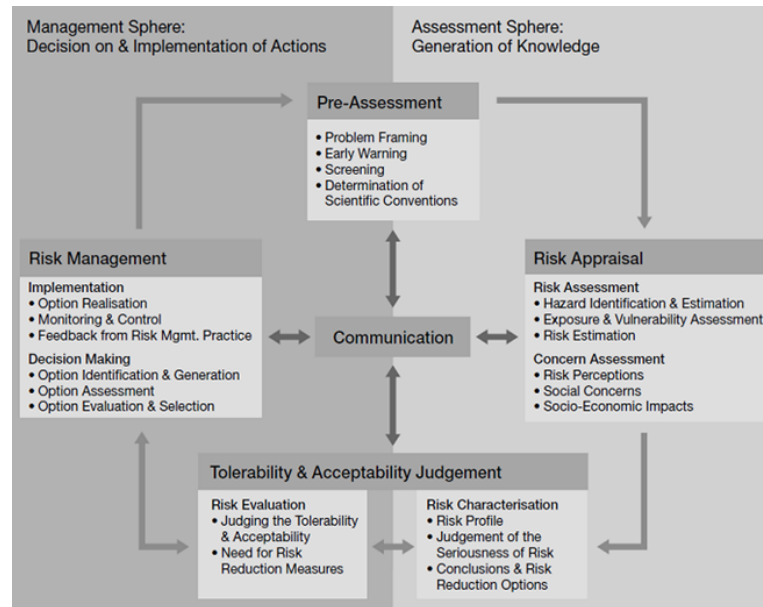


Figure 5. The risk assessment and decision-making process according to the International Risk Governance Council (IRGC, 2005 [32]) with at the right-hand side the risk appraisal part starting with the analysis and separately at the left-hand side the management/decision-making part, but all connected through communication.

As a third practical measure once a study is finished, the risk assessment (RA) report should be scrutinized and commented by an independent expert organization for a peer review, and given a second opinion on recommendations. Recently, Lathrop and Ezell [31] made a strong plea for such a step and presented a flow scheme of the assessment process, as shown in Figure 4. They also proposed tests for the various elements and described possible shortfalls. On a detail level, the above maturity model of Rae et al. [30] can serve as guidance to locate flaws. However, as we can learn from the experiences with the ‘safety case’ result assurances described in Section 2, nothing can be guaranteed; it is only that credibility can be increased.

As a final measure, the recommendations of the International Risk Governance Council (IRGC) [31] for the assessment and decision-making process of high-risk projects should be remembered and applied, see Figure 5. These require from the start good communication among all stakeholders, and within a Risk Governance program just as Lathrop and Ezell [31] based on their experience recommend, strict separation of the analyst from the decision maker to reduce bias and encourage deliberation of all relevant information.

Based on all of the above it is clear that conducting a risk assessment study by itself is only half the work, determining confidence limits and building trustworthiness cost at least the same amount of effort.

## 5. Conclusions

- a. Study results of assessing high consequence-low probability risk of process plant installations containing hazardous materials quantitatively suffers commonly from large uncertainties that are not generally being modeled and managed.
- b. Trustworthiness and appropriateness arguments can be developed according to the approach shown for ‘safety cases’ following the Goal Structuring Notation and evidential reasoning method as proposed by Nair et al. [6], although this approach cannot result in an absolute guarantee.
- c. Several methods are available to derive failure probabilities given a certain amount of evidence. Based on precursor frequency evidence and an event tree, Bayesian methods may produce the best estimates of rare event frequencies. A spectrum of alternative methods is available to tackle in particular epistemic uncertainty for deep uncertainty cases in which knowledge exists.
- d. A few practical measures are summarized to obtain more confidence in risk assessment results: For consequence models work is being continued to develop an evaluation protocol. According to the work of Rae et al. [30], the RA maturity model gives guidance on what kind of flaws can be found in a risk assessment. Further, a peer review can support confidence in a RA, while another measure to reduce confirmation bias is to separate analyst and decision maker.
- e. Performing a risk assessment is only half the work; determining limits of confidence and building trustworthiness arguments is the other half. However, result reliability of a QRA is never 1, and the actual value of the reliability in fact unknown, even in case confidence limits are provided. Yet, a risk assessment is worthwhile to identify the physical possibility of an event with large undesirable consequence, so that decision can be made and communicated whether the risk needs to be reduced and how this can be done best.

## 6. References

1. Aven, T, and Zio, E., 2012. Industrial disasters: Extreme events, extremely rare. Some reflections on the treatment of uncertainties in the assessment of the associated risks, *Process Safety and Environmental Protection* 91, 31-45.
2. Interim Defence Standard 00-56 Part 1 - Issue 5, in, UK MOD (2014) (Issue 4, 2007, <http://www.skybrary.aero/bookshelf/books/344.pdf>)
3. HSE, The Offshore Installations (Safety Case) Regulations 2005, UK Health and Safety Executive, (SCR05) <http://www.legislation.gov.uk/uksi/2005/3117/contents/made>
4. Sujan, M.A., Habli, I., Kelly, T.P., Pozzi, S., Johnson, Ch.W., 2016. Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices, *Safety Science* 84, 181–189.
5. GSN Community, 2011. GSN Community Standard No. 1, November 2011, © 2011 Origin Consulting (York) Limited, on behalf of the Contributors, [http://www.goalstructuringnotation.info/documents/GSN\\_Standard.pdf](http://www.goalstructuringnotation.info/documents/GSN_Standard.pdf).
6. Nair, S., Walkinshaw, N., Kelly, T., and de la Vara, J.L., 2014. An Evidential Reasoning Approach for Assessing Confidence in Safety Evidence, Simula Research Laboratory, Technical Report 2014-17 November, <https://www.simula.no/publications/evidential-reasoning-approach-assessing-confidence-safety-evidence>.



7. Graydon, P.J., Holloway, C.M., 2017. An investigation of proposed techniques for quantifying confidence in assurance arguments, *Safety Science* 92, 53–65.
8. Graydon, P.J., Holloway, C.M., 2016. An investigation of proposed techniques for quantifying confidence in assurance arguments. Technical Memorandum NASA/TM-2016-219195, National Aeronautics and Space Administration, Hampton, VA, USA.  
<http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20160006526.pdf> .
9. Paskan, H.J., Rogers, W.J., Mannan, M.S., 2017. Risk assessment: What is it worth? Shall we just do away with it, or can it do a better job? *Safety Science*, Article in Press,  
<http://dx.doi.org/10.1016/j.ssci.2017.01.011>.
10. Cameron, I.T., Mannan, M.S, Németh, E., Park, S., Paskan, H.J., Rogers, W.J., Seligmann, B.J., 2017. Process Hazard Analysis, Hazard Identification and Scenario Definition; Are the conventional tools sufficient, or should and can we do much better?. *Process Safety and Environmental Protection*, Article in Press, <http://dx.doi.org/10.1016/j.psep.2017.01.025>.
11. Kaplan, S., Garrick, B.J., 1981. On the quantitative definition of risk. *Risk Analysis*. 1 (1), 11–27.
12. Yang, M., Khan, F.I., and Lye, L., 2013. Precursor-based hierarchical Bayesian approach for rare event frequency estimation: A case of oil spill accidents, *Process safety and Environmental Protection*, 91, 333–342.
13. Yang, M., Khan, F.I., and Lye, L., Amyotte, P., 2015. Risk assessment of rare events. *Process safety and Environmental Protection*, 98, 102–108.
14. Khakzad, N., Khan, F.I., Paltrinieri, N., 2014. On the application of near accident data to risk analysis of major accidents, *Reliability Engineering and System Safety*, 126, 116–125.
15. Paskan, H.J. and Rogers, W.J., 2011. BBN, a Tool to Make LOPA More Effective, QRA More Transparent and Flexible, and Therefore to Make Safety More Definable!, MKOPSC 14<sup>th</sup> Annual International Symposium, October 25-27, College Station, Texas, pp. 395-408.
16. Zeng, F., Lu, M., Zong, D., 2013. Using D-S Evidence Theory to Evaluation of Confidence in Safety Case, *Journal of Theoretical and Applied Information Technology*, 47 (1), 184-189.
17. Shafer, G. (1976). *A Mathematical Theory of Evidence*, Princeton University Press.
18. Dempster, A.P., 1967. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics* 38(2): 325-339.
19. Curcurù, G., Galante, G.M., La Fata, C.M., 2013. An imprecise Fault Tree Analysis for the estimation of the Rate of Occurrence Of Failure (ROCOF), *Journal of Loss Prevention in the Process Industries* 26 1285-1292.
20. Hazelrigg, G.A., 2010, *Fundamentals of Decision Making for Engineering Design and Systems Engineering*, © Copyright 2010 by George A. Hazelrigg. ISBN 0984997601, 9780984997602  
[https://books.google.com/books/about/Fundamentals\\_of\\_Decision\\_Making\\_for\\_Engi.html?id=BdCKlwEACAAJ&redir\\_esc=y](https://books.google.com/books/about/Fundamentals_of_Decision_Making_for_Engi.html?id=BdCKlwEACAAJ&redir_esc=y).
21. Bolger, F. and Rowe, G., 2015. The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? *Risk Analysis*, 35 (1), 5-11.
22. Cooke, R.M., 2015. The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight, *Risk Analysis* 35 (1), 12-15.
23. Paskan, H.J. and Rogers, W.J., 2016. What value has risk assessment? What factors do we have to reckon with in explaining its results? MKOPSC 19<sup>th</sup> Annual International Symposium, October 25-27, College Station, Texas, pp. 554-571, in particular pp. 566-568.
24. Bolger, F. and Rowe, G., 2015. There is Data, and then there is Data: Only Experimental Evidence will Determine the Utility of Differential Weighting of Expert Judgment, *Risk Analysis* 35 (1), 21-26.
25. Cox Jr., L.A., 2012. Confronting Deep Uncertainties in Risk Analysis, *Risk Analysis* 32 (10), 1607-1629.
26. Walker, W.E., Marchau, V.A.W.J., Swanson, D., 2010. Addressing deep uncertainty using adaptive policies introduction to section 2. *Technological Forecasting and Social Change*, 77 (6): 917–923.

27. Fenton, N., and Neil, M., 2013. Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, Taylor & Francis Group, FL USA, , pages 362-364, ISBN 978-1-4398-0910-5.
28. SAPHEDRA, 2015. [www.industrialsafety-tp.org/filehandler.ashx?file=14395](http://www.industrialsafety-tp.org/filehandler.ashx?file=14395).
29. SAPHEDRA HSE, 2017. Review of consequence model evaluation protocols for major hazards under the EU SAPHEDRA platform, Simon Coldrick, Health and Safety Executive Buxton UK, RR1099, 02/17 [www.hse.gov.uk/research/rrpdf/rr1099.pdf](http://www.hse.gov.uk/research/rrpdf/rr1099.pdf)
30. Rae, A., Alexander, R., McDermid, J., 2014. Fixing the cracks in the crystal ball: a maturity model for quantitative risk assessment. Reliability Engineering and System Safety 125, 67– 81.
31. Lathrop, J., and Ezell, B., 2017. A systems approach to risk analysis validation for risk management, Safety Science, Article in Press, <http://dx.doi.org/10.1016/j.ssci.2017.04.006>.
32. IRGC, 2015. IRGC, White Paper on Risk Governance towards an Integrative Approach, International Risk Governance Council, Geneva, September 2005, [www.irgc.org](http://www.irgc.org).
33. Modarres, M., Kaminskiy, M., Krivtsov, V., Reliability Engineering and Risk Analysis, A practical guide, 2nd Ed., CRC Press, Taylor & Francis Group, 2010, pp.106-110, 114, 121, and 435-438. ISBN 978-0-8493-9247-4.

## APPENDIX

### Comparison of a conventional frequentist and the Bayesian statistical methods to determine failure frequency of, e.g., a pump based on scarce evidence

#### *Frequentist statistics:*

Suppose a stand-by pump fails  $t = 2$  times out of  $n = 10$  activations, hence by relative frequency of occurrence, mean failure chance estimator  $\hat{p} = 2/10 = 0.2$ . Assuming binomial failure behavior the  $F$ -cumulative distribution function  $p_l = \{1 + (n - t + 1) t^{-1} F_{1-\alpha/2} [2n - 2t + 2; 2t]\}^{-1}$  and  $p_u = \{1 + (n - t) \{(t + 1) F_{1-\alpha/2} [2t + 2; 2n - 2t]\}^{-1}\}^{-1}$ , can be applied to calculate the lower and upper probability values with  $1-\alpha$  is interval of, e.g., 90% and  $F[f_1; f_2]$  being the cumulative  $F$ -distribution with degrees of freedom  $f_1$  nominator, and  $f_2$  denominator, see e.g., Modarres et al. [33]. This way the 90% *confidence failure limits* are found as  $\Pr(0.037 \leq 0.2 \leq 0.51) = 0.9$ .

With more observations, the epistemic uncertainty will be reduced and the confidence interval will narrow down.

The *Bayesian* approach makes use of all information available; subjective or just an estimate, and this background knowledge is applied as a prior distribution  $h(x)$ . However, as in this case this knowledge is blank, so  $h(x)$  is represented uninformatively by a uniform prior (0, 1) distribution, which is equal to a beta( $\alpha, \beta$ ) -distribution with  $\alpha = \beta = 1$  and mean  $= \alpha/(\alpha + \beta) = 1/2$ . A beta-distribution is a continuous distribution of the probability of a component being intact or failed, while it is also conjugate to a binomial distribution. The observed data,  $t = 2$  failures out of  $n = 10$  activations, is a binomial likelihood function  $l(t|x)$ .

The Bayes theorem, which in principle is just  $P(A|B) = P(B|A) \cdot P(A)/P(B)$ , with a binomial likelihood function and a beta prior distribution produces an updated  $x$  as a posterior beta-distribution:  $f(x|t) = \frac{l(t|x) \cdot h(x)}{\int_{-\infty}^{\infty} l(t|x) \cdot h(x) dx}$ .

Leaving out the constants in the distribution equations, the nominator in the above equation can be written as proportional to:

$$\propto x^t(1-x)^{n-t}x^{\alpha-1}(1-x)^{\beta-1} = \propto x^{t+\alpha-1}(1-x)^{n-t+\beta-1}.$$

The terms at the right side of the equation represent the posterior beta-distribution. In that case, applying the equation for the mean of a beta distribution, the point Bayesian estimate is:

$p_B = (t + \alpha)/(t + \alpha + n - t + \beta)$ , or with  $\alpha$  and  $\beta = 1$ , it is  $(2+1)/(1+10+1) = 0.25$ . The 90% *Bayesian credible interval* can now be calculated making use of the cumulative beta-distribution as:  $\Pr(p < p_l) = I_{p_l}(t + 1, n - t + 1) = \frac{\alpha}{2}$  and  $\Pr(p > p_u) = I_{p_u}(t + 1, n - t + 1) = 1 - \frac{\alpha}{2}$ , in which  $I$  is the cumulative distribution function (cdf) of the beta distribution that can be calculated easily using MS Excel. This results in  $\Pr(0.078 \leq 0.25 \leq 0.47) = 0.9$ , which is narrower but still close to the frequentist confidence interval.

However, if we have background information, e.g., about similar pumps, but another brand, or from the same brand but a different population not allowed to include in the classical statistics, this can be used in the Bayesian approach. Suppose in this group 15 failed out-of-100. Assume now again a beta-distribution prior, then with a binomial likelihood function  $p_B = (2+15+1)/(1+10+100+1) = 0.16$ ; and the 90% credible interval is  $\Pr(0.10 \leq 0.16 \leq 0.21) = 0.9$ . This range is much narrower than the ones before and lies within the limits found previously.