Quantifying Risk and How It All Goes Wrong

Keith Miller, Independent Technical Safety Consultant, Norwich, UK

keithpmiller958@gmail.com

Over the course of a twenty-year study into the shortcomings with risk quantification the author has identified several fundamental errors in both the data and algorithms employed in probabilistic models. These are shown to have detrimental impacts on the perception of risk and choice of mitigations. The author argues that all forms of risk prediction (as opposed to statistics) distort understanding and distract effort away from more holistic methods, such as well reasoned arguments, for demonstrating ALARP. The psychological aspects of risk assessment have profound influences on both quantitative and qualitative assessment, because there is too much flexibility and subjectivity allowed in QRA/PRA and Risk Assessment Matrices for them to add any real value. It is concluded that, whilst UK legislation does not require quantification, or the assessment of risk tolerability, there exists a 'perfect storm' of conditions that maintain the status quo and obstructs a necessary paradigm shift for how risk should be assessed. This stalemate can only be broken by a rigorous challenge to the statistical and mathematical processes, legal interpretation and guidance.

Background

UK safety legislation is based on goal setting principles, which require risks to be reduced to As Low As Reasonably Practicable (ALARP). The demonstration of ALARP is normally achieved through a Well-Reasoned Argument (WRA) that takes account of relevant factors, such as the hazards, their causes, severity and the controls in place to prevent it, or to mitigate the consequences. There is no precise definition of the measures necessary to achieve ALARP, but the principle that action should be taken unless the costs are 'grossly disproportionate to the risks' has led to Cost Benefit Analysis (CBA), which in turn requires Quantified Risk Assessment (QRA) to become a commonly accepted process (also known as Probabilistic Risk Assessment). Various industry and regulatory guidance also implies the need to demonstrate that risks are tolerable. However, this assumes that QRA is accurate enough for this purpose and that that there is a legal requirement for it, which there is not.

After ten years of experience with QRA application and model development, the author elected to drop it entirely and use WRA for all ALARP demonstrations and Safety Cases, for the following reasons:

1 ALARP is best achieved when there is a high-quality debate between the Duty Holder and the Regulator. QRA cannot model all relevant variables, therefore simplifying the assessment, which limits the number of risk reduction options available, often increasing the cost of safety measures.

2 If a major accident occurs, the Duty Holder will almost inevitably be taken to court, and the Safety Case could be a key part of the defence. It should therefore comply with legal principles, with 'admissible' and 'relevant' evidence. QRA cannot comply with these principles.

3 QRA had not produced any safety insights, it simply put a number on things that were already understood.

4 There were serious doubts about the accuracy of the numbers produced by QRA.

The WRA approach subsequently proved to be universally successful for design, operations and material change safety cases, but it also resulted in significant cost savings, as the methodology created a more holistic understanding of the relevant variables, revealing more options for risk mitigation. It has therefore proven to be a 'win-win' outcome for the Duty Holder, Regulator, workforce and public safety.

Nevertheless, peers, managers and the Regulator demanded justification for dropping QRA. The subsequent research identified seven fundamental errors in QRA for the process industries, which are explained in this paper. However, the work also led to a broader assessment of probabilistic methods in major accident assessment, including Risk Assessment Matrices (RAMs), qualitative methods and expert judgement, which are also covered, along with recommendations on how to produce a legally admissible demonstration of ALARP. Finally, the reasons for the lack of progress are explained together with recommendations to break this deadlock.

The Problem with Modelling Rare Events

QRA is much more complicated than general statistical methods, which would normally do little more than a single stage modification of a base rate (statistical average). The problem is twofold, i) major accidents are rare events so the base rates are not statistically significant and ii) there are numerous variables affecting those base rates. The QRA solution is to simulate how the accident scenario will develop, incorporating as many of those variables as practicable, on the assumption that this will be more accurate. However, this is a complex process, requiring data to be compiled on many of the variables, together with hypothetical algorithms, simplifying assumptions and even omissions. This introduces multiple opportunities for error, which would need rigorous technical and statistical quality assurance, as rare event frequencies cannot be empirically verified. In this research, virtually all the classical errors types are found at the very starting point of this process; in the collection, interpretation and application of loss of containment data.

Sampling Criteria and Confidence Intervals

All statistical samples must conform to three basic criteria:

1 The sample data must be representative of the population being considered
2 Sampling must be random
3 The sample sizes must be large enough to be statistically significant.

Statistical significance is measured by confidence intervals, which must be small if the sample is to have credibility. However, due to the complexity of QRA models, it is not possible to calculate confidence intervals around their risk predictions, although it is possible to interrogate these uncertainties in the sampled data. It is a statistics rule of thumb that sample sizes of less than thirty are unreliable. However, QRA samples may be much smaller than this, so it is important to calculate confidence intervals for Poisson distributions down to very small sample sizes, as shown in Figure 1.

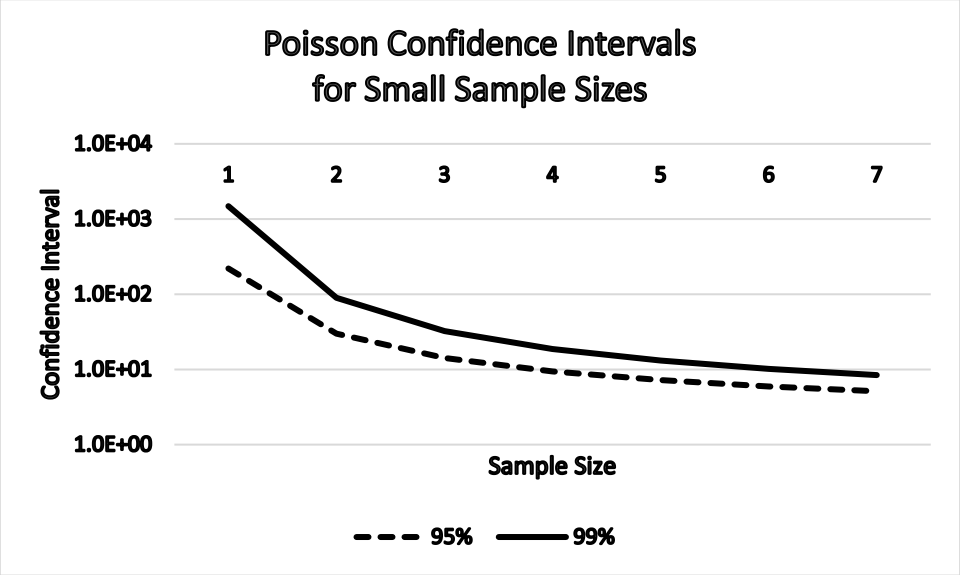

Figure 1 Confidence Intervals for Small Sample Sizes

Even the best leak databases contain gaps, with some of the most relevant categories having nought, one or two data points, with resulting uncertainties that can be measured in orders of magnitude. For example, there may be no data on 150mm holes in plate exchangers, yet some interpretations quote this frequency with an accuracy of four significant figures.

Error #1 Non-Representative Data

The leak data in the process industries does not reflect the causes of major accidents. In Table 1 the left-hand column is a list of these causes for some well-known major accidents, which are typically large, normally full bore, failures, greater than 10 kg/s. However, the right-hand column looks at the causes of the more frequent, smaller leaks, commonly known as 'weeps and seeps'. There are few, if any, major accidents that have been caused by these failure modes, yet they dominate the datasets.

Table 1 – Causes of Major Accident Losses of Containment

Major Accident Releases by Cause (small data quantities)	Non-representative Releases by Cause (vast majority of data)
Brittle Fracture (Longford)	Wear and tear on valve stem seals
Stress Corrosion Cracking (BG Rough)	Pitting corrosion
Significant impact or dropped object (Mumbai High)	Flange gasket deterioration and poor bolt tightening
Overpressure rupture (Grangemouth, Ocean Odyssey)	Poorly fitting instrument connection
Design calculation error (Flixborough, Macondo)	Passing valve
Process upset (Texas City, Buncefield, Seveso, Bhopal)	Wear and tear on door seals
Isolation/reconnection error (Piper Alpha, Pasadena)	Sampling

Because the larger leaks are not statistically significant it is normal practice to plot all leak data on a frequency vs. size graph and find the best fit curve, in the clearly erroneous belief that this overcomes the paucity of large leak data. This logic simply tries to overcome contravention of one statistical criterion (that of statistical significance) by contravening another (representativeness). By comparison, the idea that we could use data on the common cold to predict cancer rates would quickly be dismissed, even by the uninitiated.

Error #2 Defining Objectives and Aligning Data

In any statistical assessment it is good practice to define the overall objectives and relevant variables up front, so that data categories can then be specified and collected appropriately. QRA is used for many applications, often despite little evidence that the process was designed for those purposes. Examples include:

1 Average Industry Risk (to the population or individuals)
2 Installation specific risk (ditto)
3 Design, maintenance, operational or human factor risks
4 Cost Benefit Analysis (CBA) of barriers
5 Measuring or specifying Safety Critical Element performance requirements
6 Simulating the Left-Hand or Right-Hand Side of the Bow Tie
7 Installation layout arrangements

Despite this multitude of potential objectives, data collection cannot be designed accordingly, for several reasons:

1 The size, global nature (political, legislative and institutional) and diversity of the data population.
2 Data collection, algorithm development, model building, analysis and decision making involve different people and/or organisations, some of which have no formal connections.
3 Data may come from various sources. For the offshore industry it comes from five disciplines, for drilling, wellheads, topsides, risers and pipelines, each of which may have differing methods and objectives for collecting the data.
4 Available data may be based on reliability and general safety categories, rather than major accident causes.
5 The criteria for measuring leak severity, e.g. inventory loss vs. leak rate, can lead to very different results.
6 The desired categories of data may not be statistically significant.

For these reasons risk assessors have extremely limited influence over data collection and may have to compromise on categories that relate to reliability, rather than safety. The data is therefore the fixed element, thus defining the variables and forcing the algorithms to follow the structure of the data, rather than accident theory, making it incompatible with the above objectives. The following sections illustrate the potential magnitude of these shortfalls.

Error #3 The Causal Fallacy

The adage 'Correlation does not imply causation', is applicable to many of the data categories in use. Typical QRA datasets correlate leaks with pressurised equipment items, such as flanges, pipework, compressors and pumps, yet interrogation of the data reveals that almost half of major leaks do not involve the failure of any such component; which is also true of those accidents listed in Table 1. More logical correlations would be failure modes, activities, human factors, barrier effectiveness, or they could be aligned with the more advanced and established theories relating to risks of accidents, such as complexity and close coupling (Perrow, 1984), or systems analysis (Leveson, 2011). Key variables, such as design philosophy, pressure, temperature, fluid constituents and material properties cannot be incorporated. Similar problems arise with ignition probabilities, which cannot be linked to control, shutdown or electrical tripping philosophies. Data may be available on fire pump reliability, but not on the effectiveness of deluge in a real fire, and so on.

The Piper Alpha disaster provides a good example of how the equipment category for flanges can be so deceptive, because the accident involved a flange that had been left with a loosely fitting blind. But attributing this incident to the flange is illogical, because it did not fail, nor was its existence attributable to the accident. The relief valve had to be removed for maintenance, so the existence of this flange was not the issue. The acknowledged causes of the accident were Permit to Work errors, competence, shift change handover, etc.; factors that are not unique to flanges, nor equally attributable to other flanges on the installation, because they would not be dismantled for the same reasons and some may not be opened in the lifetime of the installation. Therefore, the flange is a misleading correlation that gives no clue to the real causes of the release. This is known as the 'Causal Fallacy'. Mapping either loss of containment or accidents to equipment categories works simply because they take up the entire parameter space, so correlations are inevitable, albeit meaningless.

Accident theory can be argued to have evolved through four generations. Historically, accidents were originally attributed to component failure, but this evolved to an understanding of human error, which then led to organisational/cultural issues and latterly systems analysis (Leveson, 2011). QRA therefore remains firmly rooted in first generation theory, because its data is primarily collected for reliability purposes, which closely correlates with equipment, whereas safety does not.

The historical connection with reliability data is difficult to break as it is probably the only common metric to span national and political boundaries. Global collection to more logical metrics, such as complexity or management systems, would be impractical. Nevertheless, it would be reasonable to question why equipment data is any better than simply measuring the size of the installation, either by plot area, process inventory, tonnage of steel, etc., which may be no less accurate.

The choice of data categories ultimately dictates the structure of the QRA algorithms, necessitating omissions, substitutions and assumptions, thereby distorting the user's perspective on key risk relationships. For example, installations have been built with welded joints instead of flanges to reduce QRA risks. This policy is illogical and counterintuitive, as flanges facilitate isolations, maintenance of safety critical equipment and inspection, which would make the installation safer.

Error #4 The Null Hypothesis

A null hypothesis is the argument that two samples only differ by chance. Testing the null hypothesis on various categories of equipment could therefore determine whether the sampled leak frequency for a vessel truly differs from that of a shell and tube exchanger for example, or whether this could be down to chance. If the confidence intervals of two samples overlap there will be some chance that they are the same, and the greater the overlap the greater the confidence in the null hypothesis. A pure mathematical proof is complex for very small sample sizes, especially when more than two samples are involved, but a graphical illustration is sufficient for this case.

Error #1 showed that leak databases have small sample sizes for major accident leaks and, as these cannot be related to smaller releases caused by different mechanisms, so they have very large confidence intervals, some of which are 2, 3 or more orders of magnitude for large releases, yet the range of frequencies for a given size of leak is typically less than 2 orders of magnitude. Figure 2 shows ten equally spaced items with normal distributions being ranked with 95% confidence intervals equal to the frequency range. The probability of sample #1 coming lowest in the ranking is:

$$\Pr(\#1 \text{ is lowest}) = \text{fn}\{\Pr(\#1>\#2) + \Pr(\#2<\#1), \Pr(\#1>\#3) + \Pr(\#3<\#1), \Pr(\#1>\#4) + \Pr(\#4<\#1), \text{etc.}\}$$

A rough estimate of each term in the sequence can be seen from the area of overlap between distributions, which is clearly very high and implies that the probability of correctly ordering any two items is low and for all ten items it is negligible. Each term in the sequence is smaller than the previous, but there is even a significant chance that the highest and lowest items will be incorrectly ranked.

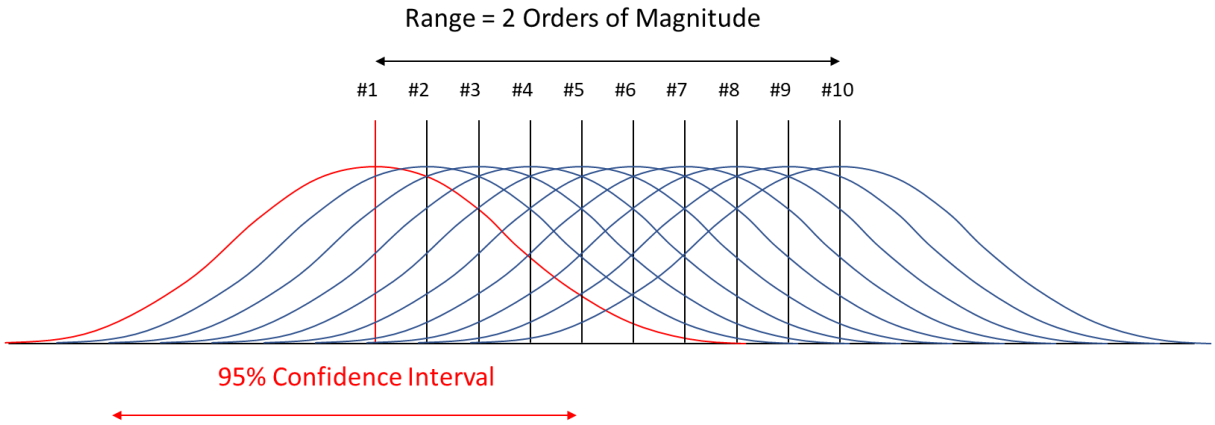

Figure 2 Probability of Errors in Ranking Order

Nevertheless, the illustration is conservative for two reasons. Firstly, a typical equipment dataset has nearly twice as many categories and secondly the confidence intervals could be much larger than illustrated (although they will vary for each equipment item). To be reasonably assured of achieving a correct ranking, the sum of all the sample uncertainties would need to be less than the ranking range divided by the number of items, which would require a vast number of major accident release samples. A pure mathematical proof would be extremely complex as would be the precise definition of the null hypothesis, given the number of items. However, it is reasonably obvious that there can be little confidence that the hypothesis would be rejected, meaning that the equipment cannot be reliably ranked and that a single frequency to represent all major accident releases would be just as reliable.

Footnote on Risk Ranking and Cost Benefit Analysis

This also illustrates another common error made by QRA analysts, who often appreciate that absolute risks are unreliable, but use the numbers for comparisons of different mitigation options, on the basis that those errors are reduced or eliminated. Firstly, assume that a Cost Benefit Analysis (CBA) is to be undertaken on options A and B for mitigating a risk and the errors in each option are independent; i.e. baseline risks are A = 5 ± 1 & B = 4 ± 1. A – B could therefore vary from 6 – 3 = 3, to 4 – 5 = – 1. So, although A & B each have error ranges of 2, the comparison has an error potential of 4. Secondly, assume that the errors are common, so A & B can be 10 times too large. A – B now ranges from 5 – 4 = 1 to 50 – 40 = 10, so the comparison error is the same as the absolute error. Given that different mitigation options almost inevitably involve different assumptions (independent errors), then CBA error potential must be larger than the absolute risk errors.

Error #5 Independence and Randomness

A key factor in any statistical analysis is the relationship between variables. When two dice are thrown the outcome on each die is random and therefore independent of the other. The probability of throwing two sixes is therefore the product of the separate probabilities (= 1/36). However, this is not true of QRA data, because failure modes may be time related and may be rectified after the first incident. Assume a corrosion life of 10 years, ±1 year. 1 item would be expected to fail after 10 years, whilst 100 of them would only reduce this to 9 years. Furthermore, failures get investigated, with repairs and/or replacements before the system is brought back into operation. The same would be true for fatigue failures and, to a lesser extent, overpressure, isolations and other activities. It would therefore be illogical to assume that an installation with five

identical process trains would have five times the risk of one with a single train. Attributing a linear relationship between risk and equipment items is therefore illogical.

Independence is often assumed in the QRA event trees, which calculate the product of various probabilities, such as detection, ignition, shut down, depressurisation, deluge initiation etc. However, these may have many common mode failures, as all items may be exposed to the same permit system, maintenance philosophy, safety culture, quality system, budgetary and resource problems, climatic exposure and even one technician's incompetence applied to different systems.

Error #6 Illegitimate Transposition of the Conditional

The conditional probability of A given that B is already satisfied, is expressed mathematically as Pr(A | B) where B is the conditional. However, unless everyday problems are expressed in mathematical notation, it is not unusual for Pr(B | A) to be mistaken for Pr(A | B), which is known as the Illegitimate Transposition of the Conditional.

One of the most researched examples of this mistake was the trial of Sally Clark, for the murders of her two children, who were later concluded to have died of cot deaths. There was no evidence of foul play, except that an expert witness, Professor Sir Roy Meadows, stated that the probability of this was only a 1 in 73,000,000. Ms. Clark was found guilty and given a life sentence.

In practice, Meadows stated the probability of two cot deaths given that she was innocent:

$$\Pr(\text{2 cot deaths} \mid \text{innocence}) = 1{:}73{,}000{,}000$$

However, this is not the question that the trial sought to establish, which was, "What was the probability that she was innocent given two cot deaths?" to which the answer was:

$$\Pr(\text{innocence} \mid \text{2 cot deaths}) = 15{:}1$$

Whether Meadows realised the difference is not known, but the error went unnoticed, and even the appeal court judges failed to understand the difference when it was explained to them eighteen months later, stating that it was "a straight mathematical calculation to anyone who knew the birth-rate over England, Scotland and Wales". It was not until the case became a cause celebre and academics wrote papers on it that the defence were able to get the judgement overturned; by which time she had spent three years in jail.

A similar problem occurs with the collection of leak data, as follows:

$$\Pr(\text{leak} \mid \text{system characteristic}) \neq \Pr(\text{system characteristic} \mid \text{leak})$$

Where a 'system characteristic' can be equipment type/item, process type, activity etc. In this case, the left-hand side of the equation is the one of interest, but the reporting systems for leaks requires Duty Holders to submit a form that describes the system characteristics, but only once a leak has occurred, which therefore reflects the right-hand equation.

Bayesian theory can be used to relate the two, i.e.:

$$\Pr(\text{leak} \mid \text{system characteristic}) = \Pr(\text{system characteristic} \mid \text{leak}) \frac{\text{The number of leaks}}{\text{The population of systems with the characteristic}}$$

However, the population values cannot be known (e.g. how often the system is started up, or how many flanges there are in that part of the process industry obliged to report leaks) because there is no reporting system for this information. The population data can be nothing more than an estimate, and a review of one such dataset revealed that it had not been updated to reflect changes in the industry over a ten-year period (Bolsover, 2013).

Another problem with the data is that it is presented as univariate relationships only, so it would not be possible to determine the leak rate on a piece of equipment during a particular activity, despite the fact that this could be a highly influential factor.

Footnote on the Difficulty of Identifying Statistical Error

An Unfalsifiable Chance Correlation (UCC) error went unnoticed in the Sally Clark case, which biased the evidence against her by a factor of 100, i.e. the 1:73,000,000 figure should have been 1:730,000. A good explanation of UCCs is given by a study (Ashwanden, 2016) that demonstrated that if people cut the fat off their meat there is a 99.7% probability that they are atheists. The point was deliberately provocative to illustrate that mathematical rigour alone is not enough. 3,200 equally absurd hypotheses were tested, most of which showed little or no correlation, although she got several high correlations purely by chance. It is therefore not good enough that mathematical protocol is followed; the background population behind the figures is just as critical, and until the correlation has been justified by a logical explanation it cannot be treated as a proof. The UCC in the Sally Clark case was because that evidence came from the CESDI report (Fleming, 2000), an epidemiological study that asked numerous questions about lifestyle and conditions before cot death and identified three variables that exhibited bias; the mother's age, family wealth and parental smoking. Each of these approximately halved the probability, so it is quite conceivable that they were nothing more than UCCs. Unfortunately, the CESDI report did not detail all the questions asked, so it is not possible to test the null hypothesis, but legal evidence should be 'beyond reasonable doubt', so there is a contextual issue that the defendant would be entitled to have only the population average used against her. In fact, the prosecution firstly modified the average by the product of all three biases, giving a factor of nearly 10, despite no evidence that they were independent of each other (even though it may be argued that wealthier people tend to be older, better educated and less likely to smoke through pregnancy). This was then rather inexplicably squared, because there were two cot deaths, therefore increasing the error to a factor of almost 100. Squaring is illogical, because it assumes that cot deaths are independent events, even though hereditary causes would be much more likely in the second child, given that they

had occurred in the first. However, despite respected academics writing papers on the case, the author has found no evidence that any of them identified this UCC. This illustrates four points:

1 mathematical rigour is insufficient without knowledge of the background data
2 the context in which the data is used is critical
3 errors may be unidentifiable without full background information
4 peer review and quality systems are no guarantee that such omissions and flaws will be identified.

Error #7 The Ecological Fallacy

An ecological fallacy is a logical fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong. In other words, the average may be greatly affected by a small part of the sample or, conversely, conclusions drawn from averages may be highly deceptive for individual sample points. This raises an epistemic conflict, where new knowledge effectively invalidates prior probabilities and makes any form of Bayesian updating impossible. Risk is epistemic (knowledge centric) because it relates to a deficit of knowledge. With full knowledge there is no risk, as the outcome would be known, e.g. a coin toss only obeys the laws of physics, but the lack of knowledge about the rotational momentum, height, air density etc. and the inability to control it means that the 'epistemic risk' is 50/50 Heads/Tails, despite that fact that it would be technically possible to calculate the outcome if full knowledge were available. Now, assume that additional information becomes available, which reveals that the coin is bent, but it is not possible to quantify the change in probability, which now could be 51/49, 99/1 or worse. So the posterior probability cannot be known, and the prior probability is no longer valid.

The coin example is analogous to accident causes. Table 1 shows that none of these causes are randomly distributed across all installations, and some do not exist on many installations. For example, brittle fracture may only be relevant to 1% of installations, so its failure data will be spread across all of them in the dataset, underestimating it by two orders of magnitude where it is relevant. Therefore, the use of generic data in QRA can be regarded as underplaying known hazards. Given the legal status of safety cases, such limitations would need to be made clear, but that would expose flaws and undermine the assessment process. Even if it were made clear the legal 'relevance' of such evidence would be questionable. Conversely, Cost Benefit Analysis (CBA) and decisions made without taking this into account would be unsound and inadmissible.

Variability of Results by Different Analysts

A study by the Risø National Institute (Lauridsen, 2002) employed six consultancies to assess the risks on a chemical installation. They were given identical information and all questions and answers were shared with all parties to ensure that there would be no bias. Figure 3 shows the results, with three of the eleven systems risks having more than four orders of magnitude variation.

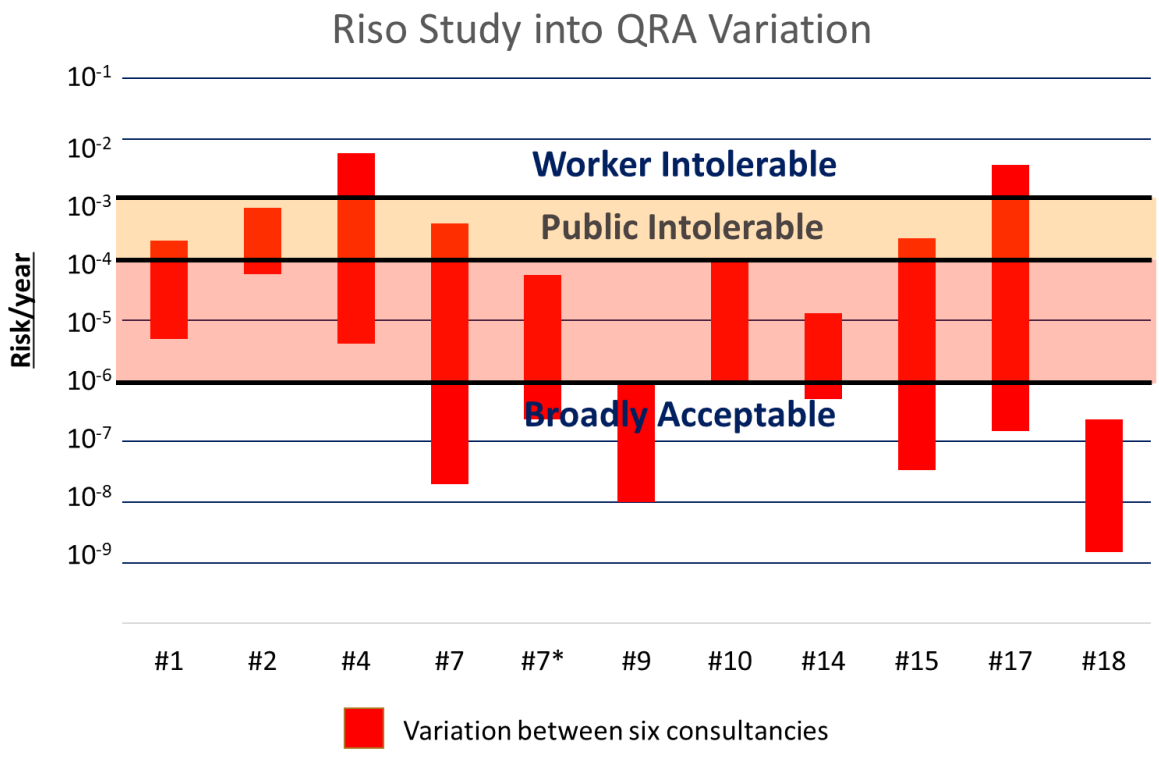

Figure 3 QRA Variation vs. ALARP Range

Mapping this variation onto the ALARP region of risk, together with the generally accepted tolerability limits for workforce and public, shows that case #17 risk predictions ranged from intolerable to broadly acceptable; the variation being up to 10

times larger than the ALARP region for workers, and 100 times larger than that for the public. Considering that CBA is intended to determine risk changes within the ALARP region, this renders the process unfit.

It should be noted that this variation is not the caused by the errors described above, which would be assumed to be consistent for all cases. A second phase of the study aligned the assumptions, reducing the variation to two orders of magnitude, except for one case that remained at four. Full alignment was not possible, due to differences in approach to the problem. The report concludes that the remaining differences were necessary assumptions caused by 'analysts guessing'. The process is therefore random within two to four orders of magnitude.

Expert Judgement

It is now becoming more widely understood that subject matter expertise relates to plausibility, not necessarily probability, and this is also recognised in legal circles, especially with respect to expert witnesses. The purpose of using expert opinion is to avoid detailing unnecessarily complex arguments behind a decision. The expert can quickly select relevant aspects of his or her knowledge, often gained over years of education and experience, to make a so called 'judgement'. The context in which such judgements are made is often to dismiss certain propositions, rather than reinforce them. For example, the expert may be able to say whether a fatigue crack is plausible, or whether the temperature could rise above a given level, without needing to explain the basis of his or her reasoning. The end users of those judgements would have sufficient confidence in the expert's reasoning to dismiss fatigue and/or temperature concerns without further questioning. However, this logic does not hold true for probabilistic assessment. Although the metallurgist would be able to judge plausibility of the fatigue crack, in the absence of statistical data on that item operating under the conditions proposed, he would not have the competence to judge the probability.

There are three basic forms of probability assessment; deductive, inductive and predictive. A coin flip and the roll of a die are deductive, because the items are symmetrical and unbiased, and the toss is effectively random, as it cannot be controlled. Inductive assessments use empirical evidence and statistics, which may be regarded as scientific, if they are properly controlled and understood, i.e. representative, random and statistically significant. Prediction excludes deduction and induction, thereby making it unscientific and raising the question of whether it can constitute anything more than uninformed guesswork.

Whilst plausibility is based on deductive and/or inductive inference, the assessment of rare event probabilities has been shown here to be a much more complex process with numerous pitfalls for the unwary. The idea that even a subject matter expert could reliably undertake mental calculations of this kind without making substantial errors is clearly unrealistic. The purpose of a case for safety is to provide a 'suitable and sufficient' demonstration of ALARP, yet the term 'prediction' has become commonplace. In the initial stages of a risk assessment the objectives are primarily to establish plausibility, e.g. HAZID and HAZOP. The difficulties arise over decisions to include or exclude risk mitigations, and this is where numerical quantification of risks becomes attractive. However, there can be no statistical data for a unique situation unless Bayesian methods can be applied effectively, which is invariably impractical because of the considerable number of variables. The Risø study showed how 'analysts guessing' were inconsistent by up to four orders of magnitude. Empirical evidence of error is otherwise difficult to obtain, because major accidents are rare events and the predictions can only be disproven if they are so wrong that a series of events occurs in the space of a few years. The Comet airliner is one such example, as it suffered total loss of the first three aircraft, but after a modification to the windows there were no more failures in fifty years of service, because engineers modified the design to make fatigue cracks implausible. No one could have judged the prior or posterior probabilities, although they may have been able to say that posterior probabilities were 'broadly acceptable', i.e. that failure was not plausible with round windows. Most deductive proof of error comes from the law courts, where gross miscarriages of justice have warranted academic challenge to the figures. The Sally Clark trial had expert witness errors of one billion times and the case of Lucia de Berk in the Dutch courts (Derksen, 2009) erred by ten orders of magnitude. The Sally Clark case was probably the most researched example but nevertheless had a UCC error of one hundred times that went unnoticed even by the academics. Humans do not have the cognitive ability to recognise enormous risk errors, despite their confidence in making such predictions.

The reason that people have such confidence in their judgement of risks may stem from their ability to judge everyday problems within an order of magnitude. For example, consider selecting a plank to walk across a ditch. Most people would regard 10mm as too thin, 100mm as safe and 1m thick enough to support a tractor. However, a calculation error of the size made in the Sally Clark case would conclude that the plank thickness should be roughly ten times the diameter of planet Earth. Errors that would be humiliatingly obvious in everyday life are easily missed when judging the risks of rare events.

These conclusions also apply to Risk Assessment Matrices (RAMs), which involve multiple parameters that are too ambiguous to add value. Risks are often described in relative terms, such as low, medium and high, which are meaningless without a reference level, e.g. driving a car could be high risk compared to taking a train, but low compared to riding a bicycle. However, such comparisons are only possible when statistically significant data is available for all options, which cannot be the case for industrial risk assessments of unique activities, procedures or equipment. There are also multiple ambiguities in these matrices, relating to their purpose, metrics (likelihood - relative or absolute), scope (plant, system, function, item or component), timescale (action, activity, number of exposures, per year or lifecycle), risk (unmitigated, top event or outcome) or consequence (historical or theoretical, most likely or worst case). They can be used by managers, operators, engineers and safety professionals with varying objectives and are easily influenced by psychological factors, such as attitudes to risk, personal agendas and external pressures, accountabilities, experience (the availability heuristic) and substitution (Kahneman, 2012), so it is not surprising that the assessments have very little repeatability. Such a predictive approach is unlikely to be more accurate than QRA, which at least has an element of logic and rigour. The matrices cannot determine tolerability of risk, because they measure one risk to the population, not all risks to one individual. Other

applications use them to determine the level of risk assessment required, i.e. proportionality, but this can be better judged on potential consequences alone.

Context is key, and whilst predictions may be acceptable in business, investments, project planning and prioritising variables for research purposes, a higher standard will be demanded in a legal context. It is therefore clear that expert judgement should never be used for probabilistic arguments. In a legal context it is not unreasonable that any probabilistic argument should be transparent and reasoned deductively (plausibility) or inductively (based on Bayesian inference). In practice this precludes risk quantification from the assessment of major accident hazards, except that some risks could be shown to be negligible, but only based on a well-reasoned argument to demonstrate that sufficient truly independent and effective barriers were in place to make the event implausible.

Legal Compliance

A review has been carried out on the following UK safety legislation:

- Health & Safety at Work Act (HASAWA)
- Control of Major Accident Hazards (COMAH)
- Offshore Safety Case Regulations (OSCR), including Prevention of Fire, Explosion, Escape and Evacuation Regulations (PFEER) and the Management and Administration Regulations (MAR)
- Nuclear Installations Act (NIA)
- Railway Safety Regulations (RSR)

The author could not find any requirement to quantify risks, rank them, or for them to be tolerable. The fundamental principle behind these regulations is that all risks should be reduced to ALARP (or So Far As is Reasonably Practicable, SFAIRP, which is effectively the same thing). The OSCR 1992 did contain a requirement for the impairment frequency of Temporary Refuges to be less than 1.10^{-3}/year and for 'suitable and sufficient QRA' to be carried out, but both were rescinded in 2005 revision of OSCR (quite possibly due, in part at least, to the Risø study). Furthermore, the author has gained regulatory approval for OSCR and COMAH safety cases that contain no risk quantification.

Although there is a significant amount of industry and Regulator guidance that refers to quantification, ranking and tolerability, this is not binding and is partly a legacy of revoked legislation.

Due to some infamous miscarriages of justice, caused by experts giving erroneous probabilistic evidence, the Royal Statistical Society have produced four Practitioner Guides for Judges, Lawyers, Forensic Scientists and Expert Witnesses (Aitken, 2009 to 2014). Relevant points from the guidance are:

1 Expert witnesses must have appropriate competence in statistics if they are to give probabilistic evidence. This is clearly not the case with QRA, given the errors exposed in this paper.

2 Acceptable methods of modifying 'base rates' are given, emphasising a form of deductive Bayesian inference, and discouraging the use of mathematical formulae. The documents only refer to one stage of modification, (from prior probabilities, known as base rates, to inferred posteriors), thereby indicating that multiple modifications, such as the complicated algorithms used in QRA, are too complex to be admissible.

3 Evidence must be relevant to the case, (which is, in any case, a requirement of the courts). Statistical data must be representative, which is not the case, and averages do not reflect known system hazards, which creates the epistemic conflict described above.

4 All assumptions must be stated, and independence must be demonstrated, never assumed. With QRA assumptions are hidden in the data interpretation, and models. Many variables are not independent, as described in error #5, as they have common mode failures.

QRA contravenes all these principles. Only a Well-Reasoned Argument can be legally admissible.

Concluding Discussion

An analyst would ideally want to know the frequency of failure for a given cause, i.e.:

$$\Pr(\text{major accident leak} \mid \text{cause})$$

However, the foregoing indicates that the result could be more like:

$$\frac{\Pr(\text{statistically insignificant chance correlation} \mid \text{nonrepresentative leak})}{\text{Dilution factor}}$$

If strict mathematical rigour had been applied, the best that could have been achieved with the data would be:

$$\Pr(\text{non-representative leak} \mid \text{process industry})$$

Which is of no identifiable value to the purpose of demonstrating ALARP. The evidence on both errors and variability leaves no doubt that QRA is unfit for measuring absolute risk or making comparisons (CBA).

Despite being non-representative the loss of containment data has been shown to be insensitive to any of the factors that cause risk deviations, so the Left-Hand Side (LHS) of the Bow Tie is essentially fixed. This might be acceptable if such limitations were generally appreciated and acknowledged, except that the meaningless equipment correlations (that cannot

be differentiated statistically because of a convincing null hypothesis), create the deceptive impression that QRA does model the LHS of the Bow Tie. Furthermore, ignition data and shutdown system reliabilities are also predominantly generic, so the early barriers on the Right-Hand Side are also inflexible. There is rarely statistically significant shutdown performance data, as these are complex systems, influenced by intangibles, such as detection variables (leak location, layout, meteorological conditions etc.), maintenance and testing regimes and the hazard itself (which can never be tested), so reliabilities must be assumed. This leads to the conclusion that the most optimistic objective of such an analysis would be to obtain an industry average for three major accident frequencies i) toxic clouds, ii) fires and iii) explosions. Any attempt to segregate these frequencies further, to reflect accident causes, ignition, equipment counts, size of installation etc. would be statistically unsound.

If there was a valid reason to produce such figures, it would be illogical for Duty Holders to calculate them anyway. The generic nature makes it a natural role for the Regulator, to publish standard figures for the process industry and to do this based on major accident statistics, which would be no less accurate than QRA.

(NB. Some causes of loss of containment, such as overpressure or brittle fracture, can be influenced by Instrumented Protective Functions (IPF), which are managed by Functional Safety Assessments (FSAs) and/or Layers Of Protection Analysis (LOPA). These methods determine the necessary reliability of control systems to prevent failure, but they must be based on the probability of the relevant deviation, e.g. temperature drop, pressure surge; not leak frequencies.)

The only remaining benefits of QRA would therefore lie on the RHS of the Bow Tie, e.g. layout, escalation potential, personnel proximities, fire-fighting, passive fire protection, blast protection, escape, mustering and emergency response. An ALARP demonstration details the barriers, their effectiveness and justifies why other barriers have not been included. To reject a barrier, such as a fire-fighting system, requires a comparison of the risk with it and without it, but there is no statistically significant data available on which to base these probabilities. A typical assumption made in QRA models is that fire-fighting systems are 50% effective. No explanation for this figure is given, which would appear to be nothing more than taking the half way point between 0 and 100%. Nevertheless, there is empirical evidence on pump reliabilities, heat outputs, radiation levels, escalation times with and without deluge, etc. There are also deductive arguments for fire scenarios, personnel proximities, escape possibilities, the effects of shutdown and depressurisation and even studies that illustrate the ineffectiveness of systems in larger fires and the potential for superheated steam to increase risks to people, etc. The 50% risk reduction factor therefore appears to be nothing more than a crude heuristic, probably based on a visceral view that fire systems must be a good thing but may not be perfect. The situation also creates an epistemic conflict, as described above. If a fire-fighting system is deemed unnecessary or undesirable, then a well-reasoned argument would need to be made for the plausible fire scenarios, based on deductive inference and empirical data.

Unlike the fire-fighting system, not all decisions are binary though. For example, consider the determination of an ALARP blast design strength for a control building. In this case exceedance curves (frequency vs. overpressure graphs) will be developed for the range of explosions possible. These will be used for Cost Benefit Analysis (CBA), to compare the cost of various design strengths with probabilities and expected loss of life. It will also be necessary to place a monetary value on a person's life to determine the maximum wall strength justifiable. Unfortunately, the exceedance analysis suffers the same data errors as above and therefore cannot reflect the loss of containment frequencies for the same reasons. It will only be sensitive to deterministic explosion modelling of different gas cloud scenarios. A more rigorous approach would be to consider credible leak scenarios by reviewing HAZOPs, HAZIDs, ignition sources, activities in the area and equipment/piping sizes to establish worst reasonably foreseeable failures, e.g. a 25mm release at the worst location, which would create a given cloud size and overpressure on the control room wall. The design case has then been established by deductive inference, giving a sound argument for what is 'reasonably practicable' which, when approved by the Regulator, creates a sound legal defence in an admissible document.

The main argument for using QRA is that the Regulator will, quite reasonably, push for a higher standard than the Duty Holder may regard as reasonably practicable. The threat of safety case rejection, and the need to appeal, is a powerful disincentive to challenge this. This causes a ratchet effect, raising the standard of good practice, which may ultimately impose disproportionate costs on industry, often due to visceral rather than technical reasons, e.g. whether to stop production during a specific activity because it a) feels right, or b) has identifiable benefits. The safety case approval process places accountability on both the Duty Holder and Regulator, but probabilistic arguments can dilute these by giving individuals the excuse that they could not be expected to understand the complex QRA models. It is therefore a convenient means of gaining agreement, provided both parties are prepared to ignore the errors involved. However, this compromise is to the Duty Holder's detriment, leaving her with a weak and inadmissible legal defence. Safety cases are not rejected lightly though and provided a sound, deductive WRA is put forward the Regulator needs a robust counter argument to refuse approval. This will raise the quality of debate, keeping the discussion at a technical level, whilst minimising the politics and visceral opinions that may otherwise undermine good decisions.

The statistician George Box famously said, 'All models are wrong, but some are useful'. From a technical perspective, QRA has no useful applications, but it might have some psychological and political benefits.

The Well-Reasoned Argument (WRA)

It is clear from the foregoing that the objective under UK law must be to demonstrate that the risks have been reduced to ALARP, using legally admissible arguments. A WRA could range from a simple self-contained justification to a more comprehensive report that incorporates Bow Ties to illustrate layers of protection, plus references to supporting studies and any separate generic text, such as management systems, where appropriate. It describes the relevant aspects of the system, what can happen, how it may be prevented, the risk mitigation in place and reasons why good practice, standards or any measures with obvious potential have been rejected.

Table 2 provides a framework for the assessment. The guideword arrangement is ordinal for many problems, but not rigid, and should be a suitable checklist for any risk assessment from changing a light bulb to designing a nuclear power station. The objective is to work through as many guidewords as necessary to produce a 'suitable and sufficient' ALARP demonstration. Many straightforward activities may be resolved using compliance with standards, good practice and company rules, but more complex systems could require technical studies. A pragmatic approach is iterative, beginning with an attempt to write the ALARP demonstration, establish the gaps in the argument and then undertake appropriate studies to provide the missing information. The need for some studies may be obvious from the outset, but this sequence helps to avoid any unnecessary work. The table may also be used as a summary, especially for justifying why specific guidewords are not relevant.

Table 2 Guidewords/Template for Demonstrating ALARP Using a WRA

	Guidewords	**Suggested Processes and/or Criteria**
1	Hazards & Failure Modes	Standard or bespoke checklists, FMEA & HAZID for materials and activities etc.
2	Severity of Foreseeable Hazards	Fatality/health potential, e.g. by judgement, Probits, Dangerous Toxicity Level (DTL), dispersion, fire, explosion modelling etc
3	Exposure and Proximities	Distances to populations, activities, other hazardous inventories and systems.
4	Foreseeable Scenarios and Outcomes	Foreseeable scenario development and timelines, including escalations, health effects & fatalities. Includes justification for excluded scenarios, regarded as unforeseeable.
5	Compliance	Referencing appropriate Standards, Good Practice Company Rules/Policy and/or "Recognized And Generally Accepted Good Engineering Practices" (RAGAGEP).
6	Root Causes	Including human factors, ergonomics, software, complexity, close coupling etc. (Alternatively, this may be dealt with as part of 7 below.)
7	**Hierarchy of Controls:**	
	a. Elimination	Hazardous agent, failure mode, activity, item, system or process.
	b. Substitution	Alternative means of achieving objectives.
	c. Prevention	HAZOP, Functional Safety Assessment (FSA), Layers of Protection Analysis (LOPA), Health Monitoring, Human Factors & Ergonomics Assessment (HFEA), operating constraints etc. as appropriate.
	d. Isolation/Segregation	HAZIDs, HAZOPs, FSA, LOPA, Site Layout Assessment, Fire & Explosion Analysis etc. as appropriate.
	e. Mitigation	Detection and Alarm, Shutdown systems, Physical barriers, Fire-fighting, Neutralisation etc.
	f. Maintenance & Inspection	Ref. FMEA, FSA, LOPA, HAZOP etc. as appropriate.
	g. Organisation	Ref. Safety Management System, Permit to Work, competencies, training, Operational Risk Assessment (ORA) etc. as appropriate.
	h. Procedures	Relevance, quality, readability, prerequisites, critical steps/stages identified with appropriate warnings etc.
	i. Personal Protective Equipment (PPE)	Relevance and types.
	j. Emergency Response, Rescue and Recovery	Ref. Emergency Procedures, external support bodies/systems, etc. as appropriate.
8	Societal Expectations	This is the final judgement of ALARP, if the decision is not clear from the above.

It is best to identify and dismiss implausible consequences and scenarios early in the process. These typically require an unforeseeable combination of independent conditions and/or failure modes, e.g. double jeopardy situations, where two completely independent rare conditions or failures occur simultaneously, or perhaps a specific wind velocity range, combined with a very large leak size pointing in certain direction and failure of one or more safety systems. Some causes may also be dismissed through HAZOP and layout studies, such as large releases due to brittle fracture, overpressure, impact etc. Boundaries set in this way should be laid out at the start of the ALARP demonstration. Statistical information, for such aspects as weather data, component reliability etcetera, may be included provided it can be shown to be representative, randomised and statistically significant and devoid of any predictive elements.

A typical decision could be whether brittle fracture should be prevented using an operating procedure or a change of material, which would be more expensive, although more effective. The effectiveness of the procedure would largely depend on the document's quality, management systems, operator competence etc; all variables that could not be statistically evaluated. QRA data is also insensitive to material type, so any form of quantification and CBA would be unsafe. Nevertheless, it may be possible to justify the procedural option as 'reasonably practicable' if a well-reasoned argument

could be made for its robustness and susceptibility to human error. In other words, all arguments are deductive and legally binding commitments, rather than numbers, providing transparency, auditability and accountability.

When dealing with multiple systems in one installation it is pragmatic to consider how to deal with the parameter space, which comprises of systems, processes, causes, barriers, activities, top events and consequences. Barriers such as emergency response are naturally dealt with individually, with minor amendments for other parameters, such as certain consequences. This may also be true of other parameters, such as brittle fracture or corrosion. The Bow Tie approach is a rigorous basis to start with, but it does not demonstrate ALARP and generates large amounts of relatively inaccessible data that needs to be dissected in a manner that avoids repetition, but also makes the final product accessible to operations, maintenance, engineering, management and the Regulator. This is a key part of the risk assessment planning (and safety case architecture), as it can pay significant dividends in cost, time and quality of communication to end users.

This approach brings together disparate studies to produce a more coherent, inclusive ALARP argument that gives a better appreciation of whether the risks are tolerable or broadly acceptable. Experience shows that this will conclude all but the most marginal cases, which should be resolved by reference to societal expectations, perhaps by a team of technical and non-technical personnel reviewing the demonstration.

Experience shows that, because it considers more variables than QRA, WRA generates more cost-effective mitigations. In some significant cases it has even led to the removal of safety systems that were considered 'good practice' but were shown to be ineffective and/or generate other risks. The removal of fire-fighting systems on offshore gas installations is a good example that would not have been possible using QRA as it cannot simulate real scenarios and would have contravened the 'Reverse ALARP' principle.

Challenge to the Status Quo

There are many reasons why QRA is still an inherent part of the licensing of hazardous industrial sites and why it has not been challenged or prohibited. In practice, there is a 'perfect storm' of political, psychological and technical conditions that prevent change, as follows:

1. The initial image of QRA is one of scientific method and credibility, because complex computer programmes process large amounts of data and calculate results to three significant figures.
2. There are few technical papers critical of QRA, and they fail to recognise the difference between error and uncertainty, thereby assuming that errors adhere to probabilistic distributions. This raises the question of whether there is sufficient statistical competence within the industry, consultancies and the Regulator.
3. There are significant commercial, political and personal conflicts of interest, that drive a quantitative approach to safety, e.g. public and government want to know numerical risks to which they are exposed.
4. Acceptance of QRA is based on Theory Induced Blindness (TIB), Overconfidence & Substitution (Kahneman, 2011); Confirmation Bias, Normalisation (Hopkins, 2012) and numerical argument preferences (Windschitl, 1996). [TIB is where an initially plausible theory is accepted, but not fully understood. As understanding improves with time, the individual uses confirmation bias to maintain his or her belief in the process, rather than challenge it, despite mounting evidence to the contrary. The existence of guidance and company procedures, only serve to reinforce this feedback loop.]
5. Risk has become implicitly linked with probability and likelihood, as opposed to deductive inference. Regulator guidance encourages the use of QRA and RAMs, if not implying their necessity.
6. QRA is a complex end to end process, involving disparate groups for data collection, interpretation, algorithm development, model construction and final interpretation. Some of these groups have differing objectives and no formal communication channels, so there is little incentive to improve. This also makes it extremely difficult for a single person to fully understand the entire process.
7. Quality assurance is difficult to undertake holistically over such a fragmented end to end process.
8. QRA dilutes the accountability of all parties involved, because of its complexity and inadmissibility in court.
9. There is a lack of clear governance, as prediction is not a core issue for any of the technical institutions, and beyond the scope of the Royal Statistical Society, leading to an absence of sufficiently detailed policy.
10. Given that process plant and nuclear power stations are built to risks measured in thousands or millions of years, the predictions will never be verified or disproven. Nor is there any no means of sense checking these risk predictions, so they will never prompt a 'Paradigm Shift' (Kuhn, 1962).

This combination of socio-political and technical factors suppresses any incentive for a paradigm shift.

Conclusions

1. Both the theoretical and empirical evidence against the probabilistic assessment of major accidents is compelling. In the process industries these errors may be larger than the prediction range of QRA models and the results cannot be verified, or sense checked. The data is insensitive to all causes of major accidents yet makes meaningless correlations that imply that it is. The belief that using QRA for comparisons (i.e. CBA) will cancel out errors is a fallacy, because comparative errors can be larger than absolute ones. Chance correlations, fundamental statistical errors, assumptions and omissions adversely influence the understanding of risk and its control.

2. History shows that neither individuals, nor so called experts, can judge rare event risks, even by nine or ten orders of magnitude. The only form of probabilistic assessment of major accident risks that may be credible are where they dismiss them as negligible, or 'broadly acceptable', based on deductive reasoning and/or multiple independent barriers, e.g. double jeopardy, but are otherwise unreliable.

3. The law courts have recognised the failings in probabilistic evidence and set out ground rules via the RSS guidance, but these principles have yet to be adopted for the licensing of major hazard industries. QRA and RAM assessments contravene these and would not be admissible in a court of law. A Well-Reasoned Argument (WRA) is therefore the only credible and legally sound means of demonstrating ALARP.

4. Prohibition of QRA would:
   - create a shift from heuristics towards deductive inference, raising the quality of WRAs
   - generate a more holistic understanding of marginal, or difficult, problem areas
   - improve accountability, transparency and auditability
   - raise the quality of debate with the Regulator
   - produce legally admissible demonstrations of ALARP
   - cut costs and improve safety.

5. A paradigm shift from QRA to WRA is overdue but will not happen by conventional means. Because the subject is unverifiable it cannot experience a 'crisis' (Kuhn, 1962), the shift can only be achieved through formal challenge, detailed error analysis, an improved understanding of motives and raising awareness across industry, the institutions and the Regulator.

References

Aitken C., Roberts, P., Jackson, G, 2009 to 2014, Practitioner Guides No's. 1 to 4, Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society.

Ashwanden, C., 2016, You Can't Trust What You Read About Nutrition, FiveThirtyEight.com

Bolsover, A., 2013, A Public Leak Frequency Dataset for Upstream and Downstream Quantitative Risk Assessment, Det Norske Veritas AS

Derksen, T., The Fabrication of Facts: The Lure of Incredible Coincidence, Neuroreport

Fleming, P., Blair P., Bacon C. and Berry J., 2000, Sudden Unexpected Deaths in Infancy, CESDI SUDI research team, Stationary Office, London

Hopkins, A., 2012, Disastrous Decisions: The Human and Organisational Causes of the Gulf of Mexico Blowout, ISBN: 9781921948770

Kahneman, D., 2011, Thinking Fast and Slow, ISBN: 9781846140556

Kuhn, T., 1962, The Structure of Scientific Revolutions, ISBN: 8601405928269

Lauridsen, K., Kozine, I., Markert, F., Amendola, A., Christou, M. and Fiori, M., 2002, Assessment of Uncertainties in Risk Analysis of Chemical Establishments, Risø National Laboratory, Roskilde, Denmark**,** Risø-R-1344(EN)

Leveson, N., 2011, Engineering a Safer World: Systems Thinking Applied to Safety, ISBN: 9780471846802

Perrow, C., 1984, Normal Accidents: Living with High Risk Technologies, ISBN: 9780691004129

Slovic, P., 1990, Social Theories of Risk: Reflections on the Psychometric Paradigm, Decision Research, Oregon 97401.

Windschitl, P. & Wells G., 1996, Measuring Psychological Uncertainty: Verbal Versus Numeric Methods, Journal of Experimental Psychology: Applied, 1996, Vol 2, No. 4, 343-364