Constructing a Legally Sound Demonstration of ALARP

Keith Miller, IMechE Safety & Reliability Working Group, Birdcage Walk, London

The methods for demonstrating ALARP have remained fundamentally unchanged for decades. However, greater understanding of legal admissibility, the limitations of predictive methods and the benefits of new analytical processes have exposed significant shortfalls and the need for change. The paper concludes that the only way to produce a legally sound demonstration of ALARP comprises a Well-Reasoned Argument and it shows how this should be structured. The work proposes practical methods of undertaking this type of analysis and the underlying principles are supported by a review of well-known major accidents.

However, this much needed change is discouraged by an almost 'perfect storm' of psychological and political barriers that need to be addressed. With this in mind, one recommendation is to industry, institutions and regulators to overhaul flawed legacy guidance, which is increasingly being recognised as a misinterpretation of UK law.

NB. The views expressed in this paper are those of the author and do not necessarily represent those of the IMechE or any other organisations to which he is affiliated.

Feedback to: keithpmiller958@gmail.com

Introduction

The goal setting nature of UK legislation leaves it open to interpretation, and the fact that major accidents are rare means that there is very little legal precedent in high consequence industries. This has enabled unscientific methods to evolve that can only be proven unsound by subjecting them to root and branch scrutiny. Remarkably, this has not happened, and a range of predictive methods have evolved almost entirely unchallenged.

The reasons for this may be largely due to the difficulty with making a logical argument that all 'reasonably practicable' measures to reduce risk have been undertaken. This would ostensibly require a great deal of analysis, which might in itself be regarded as unreasonable for many risks. It has therefore become widely accepted that some form of expert judgement is required, in order to bring the assessment to a close. Whilst 'sound engineering judgement' has many good applications in the technical domain, when it comes to rare events relating to a unique set of circumstances its use has been extended to a point well beyond the capabilities of any person or process. Expert judgement has effectively given way to guesswork, otherwise known as prediction, which has become ubiquitous at almost any level of safety analysis.

The reasons for this are largely historical, as the concept of all reasonably practicable measures was introduced without clear guidance on how it should be interpreted. We are now 70 years on from Lord Asquith's definition of the term and have the benefits of updated legislation, legal precedent, guidance from expert bodies, improved technology and analytical techniques that facilitate practical methods of demonstrating ALARP without prediction. The time may therefore be right to re-evaluate whether this evolutionary path has gone astray, summarise what is required for legal compliance and determine whether prediction has any place in the process.

Based on these developments, the author provides a set of criteria for a Well-Reasoned Argument (WRA) to demonstrate ALARP, which should be legally admissible and sound. Achieving this in a cost effective and pragmatic way requires a systematic approach based on progressive filtering of hazards and their controls. Rigorous processes, such as Systems Theoretic Process Analysis (STPA), have evolved and could play an increasing role in hazard analysis, but it must be focussed on relevant scenarios only to be cost-effective. The WRA should therefore be capable of justifying the level of detail in any ALARP demonstration, together with the use or exclusion of such methods. Whilst phenomenological, human factors and systems studies remain an essential part of the process (and constitute convenient milestones in project plans) they can adversely affect the mindset of the analyst and result in gaps in the case for safety. A more holistic 'end to end' perspective is required, which looks at interactions between hazards, scenarios and barriers. The WRA criteria are intended to overcome these problems and demonstrate that prediction is both unnecessary and undesirable.

Statistics and Prediction

A critical distinction in risk assessment is the difference between statistics and prediction. Robust statistics are based on samples that are representative of the population under investigation, randomly selected, large enough to be statistically significant and cannot constitute Unfalsifiable Chance Correlations. UCCs occur where correct statistical protocol is followed but enough samples are taken that a correlation is found by chance alone (Ashwanden, 2016). When these criteria are met mathematical laws apply, confidence bounds can be determined, and posterior probabilities can be calculated using Bayesian theory. For example, the automobile industry has a wealth of data, because there are approximately ten fatal accidents per day in the UK alone. It may therefore be realistic to determine the risks on certain types of road or with seatbelts etc. However, the same is not possible for industrial major accidents, because there are too few to calculate prior or posterior probabilities for different categories of operation, installation, process, equipment, product or organisation. Without robust statistical data, so-called judgements become no more than guesswork, i.e. prediction.

Similarly, even when robust base data is available, posterior probabilities may be nothing more than prediction. If a proposal to install brake lights in the car's rear window, in addition to those already in place, is to be assessed there is no robust data or means of doing this and the assessor must resort to prediction. Although there may be logical arguments for installing these lights, i.e. drivers can see through vehicles in front to identify whether someone ahead of them is braking, any risk reductions

will be guesswork. Until significant amounts of data on injuries and fatalities have been acquired on cars with and without such lights there will be no means of verifying direct causal relationships or calculating posterior probabilities. Indeed, even with robust data many of the relevant variables may not even be known, or collected, such as whether cars would drive closer together, thereby nullifying the benefits.

Prediction at Low Probabilities

Although probabilistic predictions are regularly made in society, for issues such as sport, politics and investments, they are just opinion and generally made in 10% steps between 0 and 100%, which may not vary greatly between different advisers. However, when assessing very low probabilities, such as nuclear meltdowns or explosions on a particular type of chemical plant, the probabilities are quoted in orders of magnitude and the error potential increases dramatically. This is not a domain where humans can make predictions without robust statistical evidence or logic, neither of which is likely to be available. A good mathematician can win at Poker in the long run, because the problem has a mathematical basis, but no such equivalent mathematical logic applies to failure rates or human factors. Rare failure rates are unlikely to have statistically significant data and, even if they do, they may be highly sensitive to other variables, such as operating environment and maintenance, which will not be included in the data.

Because rare events have little or no empirical data there is little or no opportunity to verify such predictions or disprove them. The best evidence of error potential comes from the law courts where mistakes, which have led to gross miscarriages of justice, have later been exposed by academics. The evidence given by expert witnesses in the murder cases of Sally Clark (Hill, 2002) and Luisa de Berk (Derksen, 2009) were proven to err by nine and ten orders of magnitude respectively. This evidence was based on a supposedly scientific study (Flemming, 2000) that had badly manipulated the data, which was then wrongly interpreted. Despite a number of academics writing papers on this, even they failed to notice three Unfalsifiable Chance Correlations, which weighted the evidence against Clark by a factor of 45 times. The fact that none of these errors were apparent to anyone at the time, and all took years to be uncovered, illustrates the human incapacity to sense check risk predictions even with this magnitude of error.

Probabilistic Computer Models (QRA/PRA)

The quantification of major accident risks is inherently predictive, despite the fact that these models process large amounts of data (often quoted to three or four significant figures) through complex computer programmes. Both the data and the models contain numerous assumptions, omissions and predictions that cannot be verified. The author has identified eight fundamental errors (Miller, 2018), most of which can be found in the data or algorithms of any model:

1 **Non-Representative Data**
Because major accidents are rare the models draw upon different data, which is inevitably non-representative, and they manipulate this using hypothetical algorithms. It can be shown that there is no relationship between representative and non-representative data, e.g. small leaks on process systems make up the vast bulk of the data but they have different failure modes to the large leaks that cause major accidents.

2 **Causal Fallacy**
Correlation does not equate to causation. The models generally correlate the data to a limited number of variables, such as equipment types, although very few major accidents can be attributed to equipment failure.

3 **Null Hypothesis**
It can be shown that when the non-representative data is removed from the datasets that there is no statistical difference between the remainder, which is then simply a random distribution around a single mean, e.g. there may be no statistical difference between the major accident failure rates for a pump, compressor or pipe.

4 **Ludic Fallacy**
Major accident scenarios involve multiple interdependent variables, yet these relationships cannot be defined mathematically. The models necessitate simple mathematics, which assume independence of variables, i.e. to calculate the product of probabilities, which is typically a gross oversimplification.

5 **Ecological Fallacy**
Background probability distributions (base rates) do not necessarily represent those of a specific situation. If studies have identified specific problems, which are not reflected in the data, this raises an ethical concern, as the user is knowingly using non-representative data.

6 **Illegitimate Transposition of the Determinant**
This error is caused by transposing the determinant in a statement, e.g. ‘the probability it has four legs if it is a horse’, to ‘the probability that it is a horse if it has four legs’. Whilst worldwide data for failures can generally be found, a failure rate is the number of failures divided by the number of successes, of which the denominator cannot be collected. The data therefore cannot show the absolute failure rate; only the probability of one failure type vs. another, e.g. it may reveal the probability a plant was starting up when it exploded, but not the probability of an explosion when starting up. Hill (2002) and Derksen (2009) showed how failure to understand this aberration has led to gross miscarriages of justice, caused by errors of over one and ten billion times respectively and which took years to discover.

7 **Normal Distribution Assumption**
Most rare events are not purely random, and this has significant effect on the tails of the distribution, which are typically the region of interest.

8 **Time Probability Transposition**
Many variables that are a function of time are represented as probabilities, such as dormant hazard conditions.

Common parlance only refers to uncertainties in these models, but these are errors, which are unquantifiable and can have almost unlimited magnitude. There are no valid mathematical formulae for major accidents and complex computer programmes cannot overturn the adage, 'Garbage in, garbage out'. Both the data and the QRA models are inherently flawed, unscientific, lack transparency, contain numerous assumptions and omissions and cannot be sense checked or verified, so the results are easily manipulated and highly deceptive. Furthermore, these studies lean heavily on intuitive 'Type 1' thinking (Kahneman, 2011) and confirmation bias, which facilitates acceptance of any number that either feels right or satisfies the user's purpose, because there is no incentive to disbelieve it, nor any practicable means of challenging it.

The counter-argument is that, even though these errors cannot be resolved, this is 'the best that can be achieved'. Whilst such a principle may be acceptable for business or investment decisions, it is not acceptable in a legal context.

Perceived Benefits of Using QRA Models

The QRA process is commonly regarded as a good discipline and an additional tool for identifying hazards and evaluating their effects. Whilst the incorporation of statistical and phenomenological models into a plant architecture model may seem like an ideal solution, it attempts to do too many things in one package, necessitating assumptions and over-simplification or omission of many key variables and sub-models, which makes it insensitive to many relevant study objectives.

The author has undertaken explosion studies using QRA, Monte Carlo analysis and deterministic models (based purely phenomenological modelling, i.e. Computational Fluid Dynamics). The QRA models could do little more than rank explosions from different equipment and locations. The MC analyses are also probabilistic but more focussed than QRA, so better phenomenological models are integrated into the process. They gave greater insight, such as the effect of different wind speeds and directions (note that wind data is statistically significant), by post interrogation of the models against different variables. However, the CFD studies, which worked on a limited number of initial scenarios, were manually reviewed and fine-tuned in an iterative process. This facilitated an even better understanding of the key variables and led to more cost-effective methods of mitigating the risks.

Overall the probabilistic contribution was found to be negligible, regardless of its error potential. The deterministic approach was found to be more informative and produced a superior case for safety. It was also much more useful for communicating the hazards to plant staff and the regulator. The study costs for all three methods were not significantly different, but the deterministic approach produced better insights and solutions and was therefore the most cost effective overall. On reflection these conclusions are reasonably logical, as attributing probabilities (or more correctly predictions) cannot contribute to the understanding of hazards. Although many QRA protagonists would dispute this point, they generally fail to appreciate that their analysis is basically guesswork and therefore only attributes a numerical value to an unqualified estimate, rather than proving anything meaningful.

Risk Assessment Matrices (RAMs)

For most risk assessments QRA models are too complex and time consuming, so the use of Risk Assessment Matrices (RAMs) have become commonplace. There is no standard matrix, but examples have been developed by the Regulator, engineering institutions and industry, each plotting consequences and likelihood on a matrix that can range from 3x3 up to 8x8. One problem with this approach is that the descriptors on both axes can be ambiguous, especially the one for likelihood. Terms such as remote, unlikely, likely and highly likely have meanings that can vary greatly in different contexts. It could be argued that even minor injuries that are 'likely' would be unacceptable. Risks are typically 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 walk, but low compared to climbing a mountain. However, such comparisons are only possible when statistically significant data is available for all items being compared, 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:

1. Purpose - to determine risk to the individual (tolerability), or population (loss), or how much effort is required to mitigate the risk (ALARP)?
2. Scope - plant, system, function, item, action or component?
3. Metrics - absolute or relative risks (if so relative to what)?
4. Timescale – per action, activity, number of exposures, per year or lifecycle?
5. Risk - unmitigated, top event or outcome?
6. Consequence - historical or theoretical, most likely, worst foreseeable or possible?

The matrices therefore constitute ambiguous heuristics applied to an undefined entity. The users may be Regulators, managers, operators, engineers or safety professionals, each with differing objectives, who may also be influenced by psychological factors, such as individual attitudes to risk, personal agendas and external pressures, accountabilities, experience (the availability heuristic) and substitution (Kahneman, 2011), so it is clear why the assessments have very little repeatability. It is difficult to comprehend why such an ambiguous and subjective process has ever become adopted for legal compliance and very few lawyers would contemplate using this as evidence.

Legal Requirements & Evidential Admissibility

The basis of all UK safety regulations is that all activities should, so far as is reasonably practicable, be safe; which is more commonly stated as 'the risks shall be reduced to As Low As Reasonably Practicable' (ALARP). The legal obligations may

be summarised as; identify any foreseeable hazards, assess their potential consequences and implement any 'reasonably practicable' measures to control those risks. These requirements apply to the full lifecycle of any product, activity, process or installation, including design, construction, modification, operation, maintenance, inspection, emergency response, disposal and decommissioning.

Despite widely held opinions and recommendations from non-binding guidance, there is no requirement to quantify risks, no reference to 'tolerable' or 'broadly acceptable' risks, nor any requirement for the ranking of risks. As discussed in the introduction, the origins of these misunderstandings may date back to 1949 and the misinterpretations of Lord Justice Asquith's explanation of the term 'reasonably practicable':

"Reasonably practicable is a narrower term than 'physically possible' and implies that a computation must be made in which the quantum of risk is placed in one scale and the sacrifice involved in the measures necessary for averting the risk (whether in time, trouble or money) is placed in the other and that, if it be shown that there is a great disproportion between them – the risk being insignificant in relation to the sacrifice – the person upon whom the obligation is imposed discharges the onus which is upon him."

The word 'computation' was therefore used long before the advent of commercial computing, so its meaning may have been misconstrued. Nevertheless, the probabilistic/predictive principle was adopted as a fundamental tenet, only to be reinforced by documents such as R2P2 (HSE, 2001) and its so called 'carrot diagram'. Whilst not necessarily being wrong, R2P2 is not an interpretation of the law per se, and it falls short of drawing adequate distinctions between statistics and prediction. Whilst Lord Asquith's use of the word 'computation' may have been valid for statistically significant comparisons of risk and cost, its interpretation has extended to prediction (albeit masked with a scientific image achieved through the use of computers). This has led to an abundance of guidance both from the regulator and industry bodies, which is neither practical nor credible and has not been tested in the law courts.

The legislation relates to foreseeable hazards that are essentially controllable. However, probability implies randomness, and whether any accidents, other than those entirely due to natural causes, can be classed as random is questionable.

There has been significant legal precedent based on injuries and individual fatalities, and this has taken the view that any 'foreseeable' hazard has to be identified and analysed commensurate with the potential consequences. There is little evidence that the courts judgments have been led by probabilities.

Two significant developments occurred between 2005 and 2014. Firstly, the only legislation that required risk quantification, the Offshore Safety Case Regulations, had this requirement revoked. Secondly, due to some infamous miscarriages of justice caused by expert witnesses giving erroneous probabilistic evidence, Practitioner Guides for Judges, Lawyers, Forensic Scientists and Expert Witnesses were published (Aitken, 2009 to 2014), making all but the simplest of statistics inadmissible in court. Relevant points from this 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 or RAMs, given the errors explained in this paper.

2 Methods of modifying statistical 'base rates' are limited to a form of deductive Bayesian inference, 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 it cannot be (as discussed above).

4 All assumptions must be stated, and independence must be demonstrated, never assumed. With probabilistic models the assumptions are hidden in the data, algorithms and models, and many variables are not independent.

Regardless of this, most lawyers would advise against using such complex and easily challengeable calculations in court.

The preceding sections of this paper show how probabilistic modelling of rare events, such as accident risks, contains so many inherent errors that it cannot be anything other than deceptive. The latter part of this paper gives an historical perspective, showing that many of the world's worst accidents would not have benefitted from any kind of probabilistic evaluation and, worst still, that its use was, or could have been, a contributory factor in almost all of them. The fact that such evidence is inadmissible as well makes a strong case for desisting from risk prediction, but it does not explain how to fill the void that remains; that some form of disproportion can be attributed to the cost of mitigation measures compared to the risks reduced.

There is very little legal precedent relating to risk quantification, because it mainly applies to major accidents on 'permissioned' sites, which have not had such accidents since the advent of the regulations and therefore the probabilistic elements have not been tested in court. However, precedent for smaller accidents associates reasonable practicability more closely with foreseeability, good practice, objective evidence and societal expectations. Any deviations from these will need to be supported by a fully reasoned, objective case.

In conclusion, only the simplest of ALARP decisions can be based on statistics, such as whether to travel by car or train, or whether to fund a drug that has a statistical probability of saving so many lives. Only if the data is representative and statistically significant for both options can there be any confidence in the conclusions. The level of detail in almost all industrial scenarios is too complex for them to be simulated numerically.

The only viable means of making a legally sound demonstration of ALARP is with a Well-Reasoned Argument (WRA) that takes account of these technical limitations and legal precedent.

The Well-Reasoned Argument (WRA)

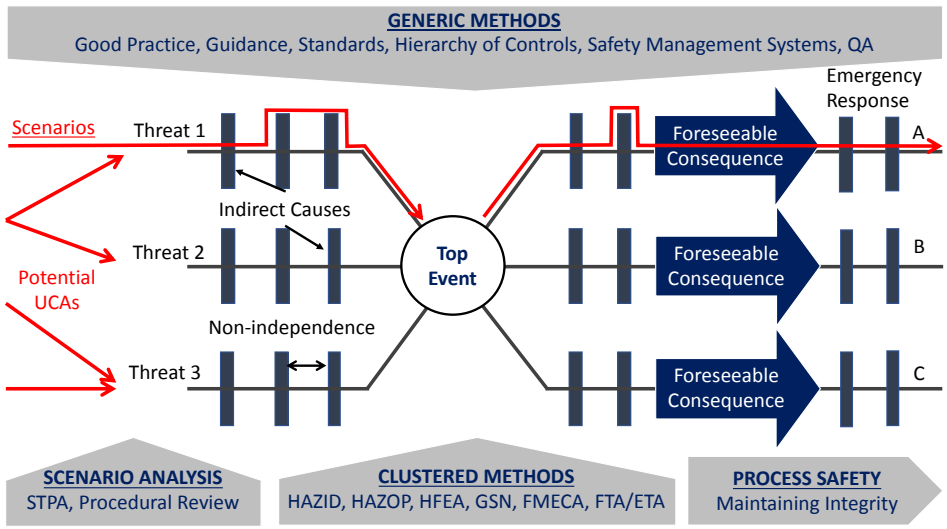
A WRA acts as a bridge between the legal requirements and the technical analyses. It is a summary of available evidence and any relevant studies to demonstrate that, so far as is reasonably practicable:

1. The foreseeable hazards have been identified for all stages of the lifecycle
2. The foreseeable consequences are understood
3. Hazard sources have been eliminated, substituted or minimized
4. The hazards, and any relevant scenarios or threats leading up to their liberation, have been avoided, isolated, contained or prevented
5. The consequences associated with the hazards have either been mitigated or eliminated
6. Each barrier, or combination thereof, is effective for its hazards, threat lines or scenarios
7. The barriers have independence, redundancy and freedom from common cause failure modes
8. The analysis is sufficiently detailed in terms of generic, clustered or scenario specific threats and barriers, (with details of any relevant procedures, activities, unsafe control actions, natural events, random failures, software errors etc).
9. Appropriate good practice, guidance and standards are complied with
10. All control measures have been applied, unless greatly disproportionate to the risk, based on logic and/or societal expectations
11. Design error has been avoided (normally by reference to QA and management systems).

Figure 1 illustrates these principles using a Bow Tie graphic, showing how individual scenarios can lead to threats and bypass assumed barriers, e.g. an Unsafe Control Action (UCA) such as an incorrect valve swing creates overpressure, loss of containment and explosion. The model shows how the demonstration methods can be categorised as Generic, Clustered, Scenario and Process Safety. This discussion is focussed on the first three, as Process Safety is adequately covered elsewhere.

Many common hazards can be adequately dealt with using generic methods, e.g. applying good practice, standards etc. With more complex situations clustered methods will be necessary but where these leave gaps it may be necessary to go one stage further and adopt scenario specific methods. For example, a clustered approach might look at overpressure of a chemical process, designing the pipework to cater for all possible pressurisation scenarios. However, this solution could be unduly expensive, so a lower pipework specification might be chosen that would not cater for some situations and these would need a more detailed scenario analysis, which would work through all foreseeable scenarios that could lead to overpressure. An advantage of the scenario approach is that it can work through human factors, software and control logic etc., instead of treating these as separate cluster studies. It would also add clarity on threat lines, where a barrier might be effective for some scenarios but not others, e.g. different sources of overpressure. The Swiss Cheese model implies that holes in barriers are random, whereas they may be 100% effective for some scenarios and completely ineffective for others. A sound demonstration of ALARP may need to recognise this.

Figure 1 ALARP Demonstration Methods



Identification

A key element of demonstrating ALARP is identification, which includes:

1 Hazards (Top Events)
- Any condition that could cause harm to people

2 Direct Causes (Threats)
- The failure mechanisms that could lead to a Top Event

3 Root causes
- Underlying conditions that increase risk

4 Appropriate barriers
- Prevention or mitigation of Top Events

5 Gaps in barriers (Indirect Causes)
- Scenarios where barriers are not relevant or partially effective

6 Design Error and Unsafe Control Actions
- Errors or omissions that could lead to a Threat or barrier deficiency

Hazard identification at the generic level would be based on experience, but where there is reason to believe that unique risks exist, they may then draw upon brainstorming methods such as HAZID, precursors from incidents (either within the business or externally) and/or bespoke hazard checklists. The identification of barrier deficiencies, design error or UCAs will depend on analytical techniques, which are discussed later (see also Figure 2 below).

Foreseeability

Legal precedent indicates that foreseeability is a key factor in determining guilt, as the courts will acquit whenever the accident could not have been foreseen, e.g. McLean v Remploy (1994), a practical joke that the courts deemed unforeseeable. A foreseeable scenario is one that would be reasonable to consider. The guidance to the Offshore Safety Case Regulations states that it is foreseeable that a helicopter would impact an offshore installation, but not an airliner. The premise here, is that a reasonable person would agree that, although the airliner scenario is possible, the combination of conditions that could lead to the accident may be dismissed.

The concept can be taken further to look at issues such as ‘double jeopardy’ and ‘failsafe’ systems where there is true independence of variables. Consider, as an example, an explosion on a chemical plant caused by a gas cloud. The worst foreseeable scenario will depend on the size and location of the cloud, which could in turn depend on a maximum leak size, particular wind strength and direction plus the failure of the shutdown system. These variables are genuinely independent of each other, which is a necessary characteristic for any double jeopardy argument. Assume that most of the piping on the plant ranges from 10 to 100mm diameter but there is some larger than this, albeit with minimal potential for failure, (e.g. it cannot be over-pressurised, has little corrosion potential and has no heavy lifting in its vicinity). Although it may still be arguable that a 200mm rupture is foreseeable, when combined with an unlikely wind condition and the random failure of the shutdown system a case can be made that the scenario is double or triple jeopardy and is therefore not reasonably foreseeable. This bounds the problem to the maximum size of explosion that can result to 100mm releases. Similarly, a prevention or mitigation measure introduced later could change the foreseeability boundaries, provided the changes maintain the independence of variables. Setting boundaries in this way may be an important part of the ALARP demonstration.

The key feature here is that all foreseeable hazardous situations have been identified and understood vis a vis the type of analysis justified. This can only be achieved using a Well-Reasoned Argument (WRA), which may need to consider scenarios if a clustered analysis is not sufficient.

Independence of Barriers

The RSS Practitioner Guides emphasise the importance of independence of variables, stating that this should always be demonstrated, never assumed. Whilst this may be a critical factor in probabilistic analyses, it has equal importance when making the double jeopardy type arguments used to justify the performance of barriers. Independence can apply at different levels; whilst two instrumented safety systems may be physically separate, they may nevertheless have common error modes, such as a common component that fails after a given time, the same flawed maintenance procedure or a poorly trained technician maintaining both and making the same calibration error. The event trees in QRA/PRA models automatically assume independence because they calculate the product of each variable’s probability (see Ludic Fallacy above). This can result in highly optimistic risk predictions. A WRA will need to demonstrate that the factors involved in any accident scenario are independent of each other at all levels if it is to be deemed unforeseeable. This may involve a description of the conditions necessary for the event, the causes, their prevention and any mitigations in place.

NB/ It should be noted that Bow Ties do not illustrate independence and may therefore give an unduly optimistic impression of safety. In practice, different barriers in the Bow Tie may actually be different parts of a single barrier or they may be inherently linked by some common failure mode. For this reason, Bow Ties are simply a graphical representation of a list of measures, the simplicity of which distorts the understanding of how the system works and therefore cannot constitute a demonstration of ALARP per se. Whilst a list of barriers may be important reference material, especially as part of a Process Safety system, putting them into a Bow Tie format can be more detrimental than beneficial.

Effectiveness of Barriers

Barriers may be used for prevention or mitigation of the consequences, but few are effective in all situations. This is illustrated by the Swiss Cheese model but with the unfortunate implication that the holes are random. In practice, the barrier may be 100% effective for some scenarios and 0% for others (as shown by the red scenario line in Figure 1). The obligation is to identify the latter scenarios and deal with them accordingly. This can lead to further investigation into whether these situations can be avoided (made unforeseeable) or mitigated in some other way. Any numerical assessment would effectively terminate the analysis, dismissing it as random, rather than investigating further.

Risk Trade-offs

It may often be the case that reducing one risk increases another, perhaps to a different group of persons. An example is that of maintaining a safety system on an offshore installation, where the technician has to travel by helicopter. Whether the risk benefits of the safety system outweigh the risks of travelling by helicopter is unlikely to be known. The safety system may benefit to a number of people but, if it were removed, the reduction of helicopter flights may lower the overall potential for loss of life even more. There is no legal requirement for acceptable risks to individuals, so decisions should be taken on overall risks to the population (unless, say, robust statistics showed that helicopter flying risks to the individual were disproportionately high, which could raise an ethical issue). Also, moving risks from workers to members of the public may not be acceptable, especially where they have not chosen to accept those hazards, or lack awareness of the dangers.

Rejection of Risk Control Measures (Reasoned Gross Disproportionality and Cost Benefit Analysis)

Occasionally, good practice, guidance, standards or identified control measures may be considered inappropriate or grossly disproportionate to the risk and a case will be needed for deviating from, or rejecting, them. Where the control measure is considered inappropriate a WRA will be required to explain why. However, Cost Benefit Analysis (CBA) has historically been used for demonstrating that a particular risk reduction measure would be grossly disproportionate to its benefits. For such a case to be credible it must relate clearly to the barrier in question and the risk being mitigated. The fundamental characteristic of CBA is an assumption of randomness, which may be reasonable when dealing with robust statistics, but cannot be relied upon otherwise.

Probabilities might be applied at the Generic, Clustered or Scenario levels. At the Generic level it has no purpose, because no foreseeable unique hazard conditions would have been identified and the ALARP case would be made without recourse to any Cluster or Scenario analysis. It is more commonly applied at the Cluster level, where a group of scenarios, say a Top Event, is being modelled. However, CBA is used to justify barriers, which may respond differently to many scenarios. A Cluster level probability would therefore be of no value and only a Scenario analysis would be able to satisfy ALARP. However, a Scenario analysis would close out the issues and not benefit from probabilistic assessment. In other words, to put a probability on an event there must be a logical reason/relationship for making such a prediction, but that would be the basis for the ALARP demonstration, in which case there is no reason to attribute a probability.

As the analysis detail increases, it moves beyond hardware issues to software. Once scientific relationships are exhausted and psychological ones are explored, credible scenarios may be identified but any probabilistic predictions can be nothing more than guesswork. Human error may be a function of ergonomics, layout, colour coding, familiarity, consistency with other parallel systems, competency, lack of or conflicting information etc. Human Error Probability (HEP) data cannot be representative for such precise situations and is therefore applied at the Cluster level. A sound case for ALARP must identify and assess the consequences of all safety critical decisions and actions. However, the solutions at this level of detail are normally quite straightforward error correction or minor modifications that are expedited regardless of probability.

The accident at Esso's Longford plant in Australia can be used to illustrate the point. In this case operational errors led to poor temperature control, brittle fracture, an explosion and two fatalities. In making the case for safety, it may have been necessary to demonstrate that using more expensive materials, which would make brittle fracture unforeseeable, would be grossly disproportionate, compared to relying on operational controls, such as procedures and staff competence. The use of the better material provides certainty, but the probability of operational controls failing is not nearly as clear.

The operational controls could have been examined to identify potential errors that could lead to brittle fracture, with each given a probability of failure, but HEPs would reflect average/base failure rates, not those unique to the Longford process system. A Scenario analysis must therefore be undertaken. The ethics of using HEPs which are non-representative of the Longford system ignores the next level of detail. In this case a pump had come off line for complex reasons and the exchanger was kept on line. When the pump was brought back on line it created temperature differentials that caused the fracture. There were many decisions made during the event and each in turn depended on who was making them, whether they were competent to do so and what information was available to them, to say nothing of the economic and time pressures that may have existed. There may also have been other scenarios that could have led to the same fracture event. This was clearly a sociotechnical system that is too complicated for HEPs, prediction or CBA. The only way to demonstrate ALARP for the procedural option would have been to show that there were sufficient independent barriers in place to make it virtually unforeseeable that mistakes could have been made. A rigorous procedural review or an STPA would be necessary. This could identify failure modes together with hardware and/or software methods of overcoming them, such as failsafe/redundant instrumented controls, alarms and automatic shutdown systems. A WRA should then be able to argue that brittle fracture would not be reasonably foreseeable.

Probably the best publicly documented example of CBA in decision making relates to the Ladbroke Grove rail disaster in 1999, where 31 people were killed and over 250 injured, following a Signal Passed At Danger (SPAD). Automatic Train Protection (ATP), which should have prevented the accident, had been rejected, partly because the CBA concluded that the

cost/benefit ratio was too high at 3.1/1. This was reviewed by the inquiry (Lord Cullen, 2001), which concluded that it was sound, but they did not appear to challenge some of the critical issues, omissions or hidden assumptions (or even whether a ratio of 3.1 to 1 could be considered greatly disproportionate).

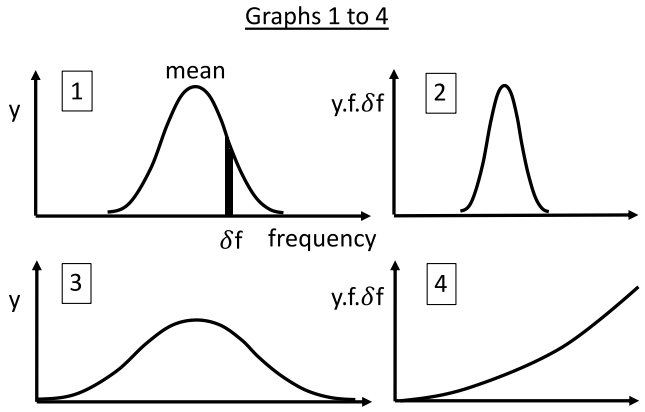
The QRA extrapolated data from minor incidents to major ones despite the possibility that some variables may drastically change for these situations. There may be non-linear relationships with variables such as train speed, train or passenger densities (e.g. during rush hour), ergonomics (signal and oncoming train visibility), weather conditions, crash resistance and fatalities (e.g. standing passengers at rush hour), etc. Foreseeable combinations of these may create disproportionately high-risk situations (which is often the very reason major accidents occur). Independence cannot be assumed; for example, the train speed may affect SPAD frequency. The models modify average data with those factors 'believed' to be appropriate to a specific situation. Furthermore, as more detailed analysis interrogates causal factors it comes to a point where the variables may be foreseeable but cannot be quantified. Three of those relevant to this accident were i) the signal was on a curve making it difficult to determine which track it related to, ii) it was only 100m after a low bridge that obscured it until the train was very close and iii) electrification of another track had also partly obscured the signal. None of these factors can be modelled verifiably and any attempt to quantify the risks would be extraordinarily speculative.

The Comet airliner showed how a small design change could affect risks by many orders of magnitude (Miller, 2018). The first three aircraft were lost making failure a virtual certainty, but modifying the windows prevented any further accidents of this kind in over sixty years of service. This shows that one small factor can totally dominate the risk picture, so the use of average data to represent specific situations may be highly deceptive. In the same way the known hazards associated with the Ladbroke Grove signal were un-quantifiable, even within many orders of magnitude. They could have increased the risks by anything from a few percent to billions of times larger. Any data is inevitably insufficient and/or non-representative, and no one can have the relevant experience, for making judgements of this type. The Sally Clark and Luisa de Berk examples showed how errors over nine orders of magnitude took years to identify, and only then because it was highly likely that innocent people had been imprisoned for life. A controlled study sponsored by the HSE revealed four orders of magnitude variation in risk predictions by different consultants (Lauridsen, 2002). The engineering community fail to recognise the significance of these errors and the Cullen Inquiry were clearly not advised properly. The cost/benefit ratio of 3.1/1 is trivial compared to the error potential and should therefore have been dismissed as inconclusive. Despite this, the inquiry report stated '*I describe the method used by W S Atkins and discuss a number of criticisms that were made of their report. I conclude that, while there was force in some of them, they were not fundamental points and did not materially affect the overall conclusion*'. Had Cullen been aware of these issues, and the eight fundamental errors listed previously in this paper, he should have come to a very different conclusion. Anyone appreciating this must question the ethics of using average and/or non-representative data to simulate a group of essentially different situations.

Furthermore, risk aversion does not appear to have been considered. The value of a statistical life lost was taken as £2.45 million, but risk aversion guidance can multiply this by up to ten times for each order of magnitude increase in fatalities, which would have completely reversed the decision.

Regardless of these arguments, a higher-level perspective reveals a more obvious conclusion. As ATP was costed at £8.9 million only 3.6 lives would need to be saved to justify the system. Within two years of the CBA report 31 lives were lost, justifying £76 million investment, ignoring the 250 injuries, train replacement costs and loss of revenue. Whilst this could be an Unfalsifiable Chance Correlation it should nevertheless have raised suspicion that the CBA was dreadfully wrong.

In practice, there were many other factors, like ATP reliability, and TPWS (the alternative system that was later chosen), which provided many of the legally sound arguments necessary for the WRA aspects of the case. The QRA/CBA was therefore at best a waste of time and at worst a constraint on any detailed analysis of the hazards. The fact that all QRA requirements were subsequently removed from any UK legislation (2005) and the RSS practitioner guides have also been issued (2011 to 2014) means that CBA of this kind should no longer be legally admissible.



The counter argument to this is that there must be foreseeable situations where potential risk mitigation measures are too expensive to be justified and some criteria for their rejection is therefore necessary. Given the error potential and need for 'great disproportion', CBA ratios would have to be many orders of magnitude before they could be credible. However, by this time the arguments behind the analysis would be self-evident and the decision could be made solely on the basis of 'Reasoned Gross Disproportionality', which is legally admissible. Truly unreasonable proposals for risk mitigation would be

equally clear to a court of law as grossly disproportionate, e.g. building a nuclear power station deep enough underground to make it safe would be grossly disproportionate to most reasonable people. Furthermore, the exclusion of prediction from an analysis incentivises the analyst to drill deeper into the arguments, and experience shows that this often provides a more complete justification for gross disproportionality. The basis for such a WRA could therefore be described as 'Societal Values and Reasoned Gross Disproportionality'.

It should also be noted that a mathematically precise CBA should not use the frequency estimate (mean), but should integrate across the probability distribution, i.e. $\int y.f.df$, as shown in Graph 1, with the point weightings ($y.f.\delta f$) shown in Graph 2. With robust data $y.f.\delta f$ peaks close to the mean, but as the uncertainty or errors increase, (e.g. the 90% confidence figure is more than an order of magnitude above the mean), then $y.f.\delta f$ continues to increase and tends towards infinity, as shown in Graphs 3 and 4. The data points away from the mean would therefore have greater weighting than the mean because f increases more than y decreases. Therefore, once the errors exceed an order of magnitude, CBA tends towards zero, making any proposed risk mitigation mandatory. The preceding evidence shows that all QRA uncertainties and errors are inevitably greater than an order of magnitude, therefore proving CBA to be an unsound method of rejecting any risk control measure.

The author has used WRA successfully for twenty years. Four noteworthy examples of how considered good practice was challenged to reject expensive risk mitigation measures were:

1. the shutting down offshore gas platforms if the fire-fighting system fails
2. shutting down when moving mobile rigs alongside gas platforms
3. the installation of radar on offshore installations to alert to potential ship collisions
4. the building of a second, bridge linked accommodation platform to separate personnel from hydrocarbon risks.

Each of these challenged considered good practice, but the WRA yielded unexpected benefits, revealing previously unrealised hazards. The analysis of fire-fighting systems questioned their mitigation foreseeability and identified risk trade-offs. QRA models normally assume an arbitrary risk reduction for fire-fighting systems, but the analysis revealed that it was incapable of preventing escalations and would produce large quantities of superheated steam, conversely increasing risk. Maintenance and testing of these systems exposed technicians to risks from helicopter flights and other platform hazards.

The process of analysing rig moves alongside installations involved detailed scenario analysis that showed there were no foreseeable consequences that could be mitigated by shutting the systems down in advance.

The additional radar systems were rejected on the basis of non-independence of barriers. QRA models for ship collision necessarily assume randomness, which ignores critical human response, such as course correction, so the predictions were fundamentally flawed. Detection and response are sociotechnical processes and splitting the responsibilities between different groups would dilute and fragment accountability, rather than creating redundancy, potentially leaving gaps in coverage. The WRA demonstrated that two barriers are not necessarily better than one, unless they can be shown to be independent.

The single platform option for the offshore installation was evaluated deterministically, using Computational Fluid Dynamics, which resulted in aerodynamic modifications to the structure and some novel changes to design philosophy, providing effective protection against all reasonably foreseeable outcomes.

In each case the exclusion of any predictive assessment led to deeper analysis and identified previously unrealised aspects that provided the case for gross disproportionality. Another advantage was the way that it raised the quality of debate between the regulator and duty holder. Because it was completely transparent, the regulator became more involved, which led to further improvements, which were often simple changes to management systems or procedures. It also had the advantage that their approval of the WRA reinforced the legal case.

CBA can only be used with robust statistical data where there are no known hazards that differ from the average. For unique or rare events CBA evidence is no more than prediction with unquantifiable error, which distracts the legal duty holders from exploring other practical solutions. If a risk control measure cannot be shown to be grossly disproportionate based on societal expectations and sound reason (i.e. robust statistics, foreseeability, independence and effectiveness of barriers and conditions, consequence reduction and risk trade-offs), then it should be implemented.

Managing Changes in Risk or New Knowledge About Hazards and Their Control

In the operational phase of the lifecycle, audits, reviews or studies may make recommendations that cannot be implemented immediately. This raises the question of whether operations can continue until the change has been made, as the costs of this may be considered to be grossly disproportionate. This partly depends on whether it would have been reasonably practicable to identify the problem earlier. If previous audits or studies had been proficiently expedited and had failed to identify the problem or mitigation measures, then it may be reasonable to say that ALARP had been maintained prior to these recommendations. Whether it could continue to be ALARP and what a reasonably practicable implementation period would be, depends on the consequence or relative likelihood reduction and any interim control measures (e.g. physical, procedural or awareness)? This will require a WRA that is consistent with the criteria laid out above. The new information/recommendations may invalidate any applicable safety case or regulatory approval, so re-approval may be necessary, but this should nevertheless strengthen the legal position.

The same principles apply to failures of barriers or changes in their effectiveness. It may be prudent to put contingency plans in place for any foreseeable failure types, including interim control measures and allowable outage times. Including these in any applicable safety case would also be prudent.

Balancing Business and Legal Decisions on Safety

Businesses often need to make decisions about ongoing risk reduction and how to prioritise safety studies or the implementation of their recommendations. Whilst CBA is often used for making business decisions it can rarely constitute a legal case and should not be considered as such. Prioritising or delaying ALARP studies and/or risk control measures cannot be justified in this way, so this would be purely a business decision, weighing the cost savings against the risk of prosecution. However, in the event of new knowledge or changes to existing risks there may be a case for delaying implementation.

A Review of Major Accidents and Assessment Methods

This section reviews the foregoing ideas using some well-known major accidents in the UK and worldwide, drawing from the nuclear, chemical, oil and gas and rail industries. Its purpose is to assess how ALARP could best have been achieved with the benefit of hindsight. Accident reports rarely provide detail on the analysis methods, but general conclusions can be drawn. Table 1 asks the following questions of these accidents:

1 Were the risks ALARP at the time of the accident?

2 Could the events be considered to be random?

3 How might the accidents have been prevented?

4 What might have been the best approach to achieve ALARP in retrospect?

5 Regardless of whether QRA or RAMs were used, could they have identified the errors or control measures, or would it have justified those that should have been considered?

With the benefit of hindsight, the systems could have been made inherently safe by designing out the problem, i.e. engineer the system to withstand all foreseeable operating conditions, or operational/procedural controls.

The following is a coarse review of these accidents, for the purposes of answering the questions above:

Table 1 A Selection of Well-Known Major Accidents

	ALARP ?	Random Equipment Failure?	Hindsight Solutions to Achieve ALARP	Analytic Method	Prediction Benefits? (QRA/RAM)
Chernobyl	No	No	Design?	Scenario	None
Fukushima	No	No	Design	Generic	Disadvantage
3 Mile Island	No	Partial	Design	Scenario	Disadvantage
Bhopal	No	No failure	Operations/Maintenance	Process Safety	None
Seveso	No	No failure	Design	Scenario	None
Longford	No	No	Design	Scenario	Disadvantage
Macondo	No	No	Design	Generic	Disadvantage
Texas City	No	Partial	Operations/Maintenance	Process Safety	Disadvantage
Mumbai High	?	?	Design	Generic/Scenario	Disadvantage
Piper Alpha	No	No failure	Procedural	Clustered	Disadvantage
Flixborough	No	No	Design	Generic	Disadvantage
Buncefield	No	Partial	Design or Procedural	Scenario	Disadvantage
Ladbroke Grove	No	No failure	Design	Scenario	Disadvantage

Chernobyl was clearly not ALARP or random, as the plant was being operated with the trips deliberately overridden. This problem was an operational one, which could not benefit from any kind of probabilistic assessment. A UCA such as this is foreseeable, and it may have been possible to design this out but would have required a scenario assessment to achieve this.

Fukushima was clearly a design failure, with a foreseeable hazard that could have been prevented by complying with good practice, as the sea wall was not high enough and the pumps were not truly independent for this scenario. Probabilistic analysis almost certainly played a part in making the wrong decision. Earthquakes do not follow normal statistical distributions (Buchanan, 2001), so the hazard should have been designed out, by specifying the largest foreseeable wave.

3 Mile Island was clearly not ALARP, as the valve position indicator only measured power to the valve and not its position. Whilst the valve failure may be regarded as technical and random the cause also involved UCAs. However, the decision to install a power only valve indicator would have been through a lack of understanding of its criticality. QRA could have easily argued that the probability of the valve being activated, then failing and the position indicator failing as well would be well below 'Broadly Acceptable', because it would assume a low failure rate for the indicator. In order to ensure the design

specification is correct there needs to be a link with the analysis. Whilst clustered analysis (HAZOP/FMECA) may have been capable of identifying this it was almost certainly undertaken and failed, so a scenario analysis should have been more reliable.

Bhopal was a case where a reactive process was not being operated within ALARP, with poor maintenance, procedures and safety culture. There were some design errors, but not significant enough to have changed the outcome greatly. The lack of a proper Process Safety management system was almost certainly the key factor. QRA could not identify or resolve any of these issues.

Seveso was a reactive process like Bhopal, which failed for very different sociotechnical reasons. The plant was being shut down, which created significant and unmonitored temperature changes due to design and operational deficiencies. Clustered analysis, such as HAZOP, appears to have failed to identify these issues, so a scenario analysis may have been the only solution. QRA could not identify or resolve any of these issues.

Longford has been discussed above, where it was shown how complex sociotechnical issues were not appreciated, leading to brittle fracture. The problem could have been designed out, based on cluster analysis, but the decision to reject the inherently safe option justified scenario analysis. QRA and CBA have been shown to be incapable of assessing these options and if used would likely have led to the wrong decision.

Macondo failed, primarily due to the use of an experimental cement that could not contain the pressure. A second barrier, the Blow Out Preventer, was only 90% effective as it could only cut the pipe, not the connectors. On the face of things, the issue was a relatively straightforward technical failure. Retrospective bow tie analyses have led to confusion, with opinions varying as to whether there were less than two or up to six barriers (Hopkins, 2012), as industrywide definitions for barriers remain ambiguous and inconsistent. There is extensive good practice available on well design and this was not followed so the failure was at the generic level. The setting of robust goals and auditing them may have made the difference, as two ineffective barriers would not have been acceptable. There was clearly a Quality Assurance problem here, which may not have been resolved even with a good Process Safety system. QRA would have only distorted understanding as it cannot know the risks of an experimental barrier.

Texas City was the event that led to the development of Process Safety systems (Baker, 2007), which could be applicable to many types of commercial pursuit. Whilst there were potential design improvements, the event was essentially operations and maintenance related. HAZOP had the potential to identify and resolve the problems, but the procedural barrier failed due to root causes. The accident led to a change of thinking with QRA, because temporary buildings were justified using the probabilistic approach. Prediction methods were subsequently removed from API 753.

Mumbai High was a ship collision with a riser that caused fires and escalations, killing 20 people and destroying the installation. This is probably the nearest example to a random accident, as the ship hit the riser immediately creating a full-blown disaster scenario. However, locating the riser where vessels offload was ostensibly illogical. Whilst the initiating collision might be regarded as random, the outcome was the result of flawed decisions during design. Placing a riser away from the location where ships are offloaded, or protecting it in some other way, is considered good practice. It is reasonable to assume that the decision was based on some form of QRA/CBA. This event alone cannot prove that this was not ALARP, but it is more than likely that this was the case. Given that there was sound reason for not locating the riser elsewhere, or that protection was not reasonably practicable, then the sociotechnical and close coupling aspects of this hazard would justify scenario analysis of offloading operations.

Piper Alpha was an operational accident that was exacerbated by poor mitigation. The accident was foreseeable, prevention was a management systems issue, whilst mitigation was predominantly design. Amongst other things, the system had human error/ergonomic implications because the relevant relief valve was not on the pump skid, so it was not obvious that the skid was unavailable when it was removed. This, together with the permit to work system and isolation procedure, made this a sociotechnical problem that was not picked up through clustered methods (assuming they were undertaken) may only have been identified with a scenario analysis. Prevention could therefore have been technical or operational. Mitigation was primarily through generic improvements to prevent escalations and provide a safe place for personnel. QRA was not commonplace when Piper Alpha was designed, but it would only have diluted these issues and distorted understanding of the hazards.

Flixborough was perhaps the simplest accident to analyse, because it involved a catastrophic failure of a replacement bellows that immediately led to a full-blown disaster. The cause was a generic level failure of Quality Assurance and Management of Change.

Buncefield was not technically a major accident, because there were no deaths, but the fires and explosions were extensive. The direct cause was a random failure of an instrumented alarm that would have tripped the pumps filling the tank, or alerted operators to the problem. The failure was a foreseeable event that could have led to multiple fatalities, although the presence of people may have led to earlier intervention and possible mitigation. The scenario led to overflow of the tank bunds, which QRA or RAM might have regarded as a ‘broadly acceptable’ risk, or unforeseeable, given trip failure and no personnel intervention. However, it was foreseeable that filling would happen on a Sunday with no personnel present, so a scenario-based approach, such as a detailed review of the filling procedure, should have revealed the need for additional barriers or failsafe systems.

Ladbroke Grove rail was a rail disaster caused by a signal passed at danger (SPAD), which has been discussed in detail above. The QRA/CBA work for this has been shown to be fundamentally flawed and the accident was foreseen. This was a case with known design hazards and UCAs, that clearly justified scenario analysis.

Concluding Remarks to the Historical Review

1 None of the situations were considered to be ALARP at the time of the accidents.

2 Very few of the accidents could be regarded as random equipment failures, the risks were controllable, and most could have been prevented by design changes.

3 Prediction and CBA, whether based on QRA or RAMs, serves no useful purpose and could, or would, have been a major adverse factor that led to many of the accidents.

4 Bow Tie analysis would not appear to have benefited the analysis of these accidents. Whilst there is value in listing the barriers, which is a necessity for Process Safety systems, and it may act as an indirect hazard identification process by stimulating brainstorming, it tends to act more as a justification for safety, because barrier counting, whether conscious or unconscious, may be the prevalent influence.

5 The analysis would appear to support the concept of addressing generic issues first and moving to clustered and scenario methods in a systematic manner.

6 The majority of the sample accidents could have been designed out and involved Unsafe Control Actions (UCAs) in non-random, foreseeable scenarios. This implies that wherever inherent safety is substituted with control measures there should be greater emphasis on scenario-based analysis.

7 The analysis implies that scenario methods were underutilised, as cluster methods have failed in many instances. This is not to say that cluster methods are inadequate, as they may have prevented many major accidents, so they remain appropriate in most cases. However, it is clear that a systematic process to filter hazards and identify where scenario methods are justified would be a significant step forward in safety analysis.

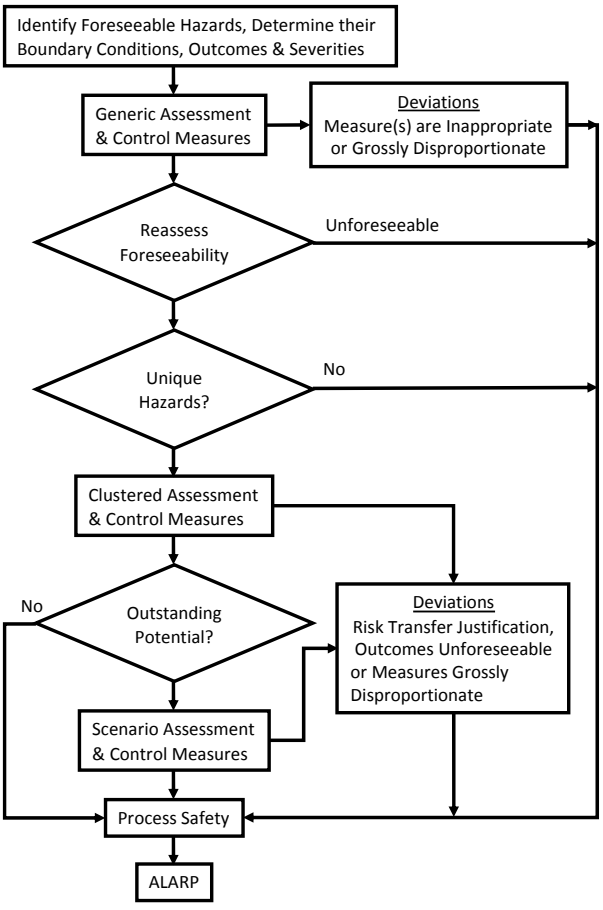
Proposals for a Pragmatic Approach to ALARP

This section looks at applying the foregoing principles into a framework that has the potential to address the lessons from the historical review. Figure 1 established the principle of categorising the ALARP methods into Generic, Clustered Scenario and Process Safety and Figure 2 suggests a means of putting these together in a decision flowchart that progressively filters out hazards and scenarios as adequate solutions are found.

Conventional methods, such as HAZOP may have failed to prevent incidents such as Longford and Seveso, so a means of progressing the analysis to another stage is needed. The 'Outstanding Potential?' decision point needs to identify both known and unknown scenarios.

For example, where barriers cluster hazards, e.g. a relief valve could protect from overpressure from a number of different scenarios, there may be no outstanding scenarios and therefore no need to do any scenario assessment. However, if it is not designed for all scenarios, or the system is not 100% reliable, there may be a need for additional barriers, and these could relate to the individual over-pressure scenarios. The Longford scenario was a complex situation that could cause brittle fracture, but it should have been apparent that clustering was not reliable and specific additional barriers needed to be included in the pump failure scenario, to name only one. These could be hardware or software barriers that could only be established from more detailed scenario analysis.

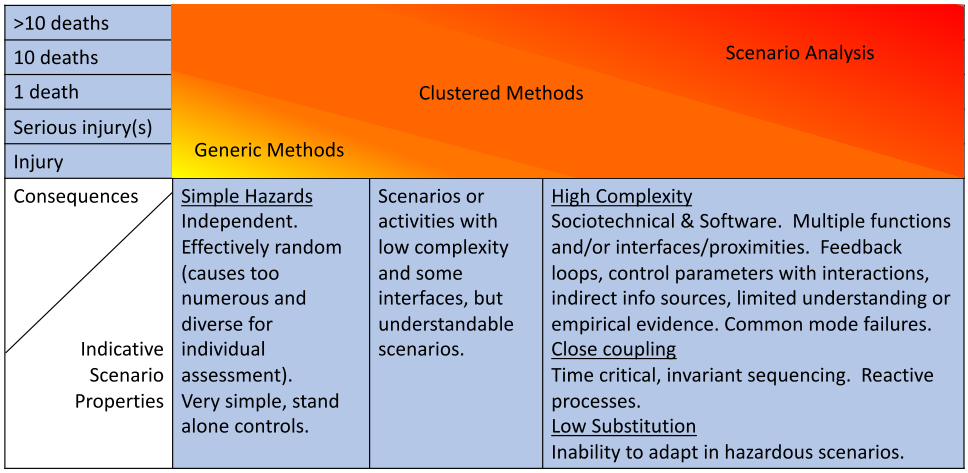
Figure 2 ALARP Demonstration Process



The Seveso scenario could be argued to be less knowable and HAZOP may have failed to identify whether further instrumentation was necessary to cater for other scenarios. It could therefore have been necessary to recognise that the system and hazard characteristics indicated the need for further analysis. Figure 3, is intended to show how the key qualities of a system or scenario may relate to the analytical effort that would be justified.

This draws upon cross-industry research to find common factors for risk (Perrow, 1984), which established Complexity, Close Coupling and Substitution as critical factors. Seveso would score highly on these as it was a reasonably complex system that was close coupled (reactive process) that had low substitution, i.e. not many alternative barriers that could have been brought into effect once the problem arose. Whilst the descriptors are not entirely objective, they provide a better means of determining the proportionate level of analysis than probabilistic models and RAM assessments.

Figure 3 Proposed Proportionality Matrix



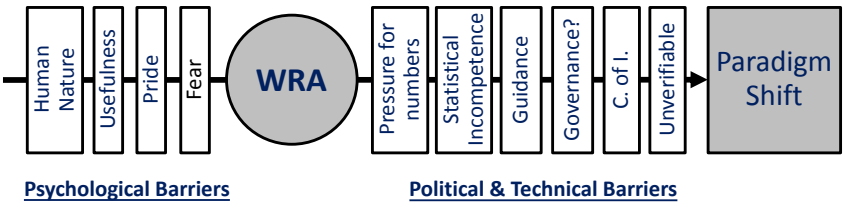
The key strategy with any WRA is to continually check for gaps in the argument, drilling deeper until they can be closed out. However, ALARP is not about arguing that something is safe; rather demonstrating that all reasons that it may be unsafe have been identified and all reasonably practicable measures to control those risks have been implemented.

Changing the Status Quo

Despite all of the legal and technical reasons why prediction should be replaced by more objective processes, it has prevailed for decades. Because the process cannot be verified empirically it constitutes pseudoscience operating in a technical environment. There is an almost 'perfect storm' of psychological, political and technical reasons why status quo remains, which is illustrated in a Bow Tie format in Figure 4.

The main reason why prediction has not been properly challenged can be explained by the way humans make decisions (Kahneman, 2011). This divides decision making into two types, 'Fast Thinking – Type 1' and 'Slow Thinking – Type 2'. Ideally all decisions would be Type 2, which is a systematic, deliberate, logical evaluation of all the facts. The problem is that this is too slow for the multitude of every day decisions, which rely on a more subconscious method of pattern recognition (Type 1). For example, if a dentist wore a fireman's jacket the patient would instinctively realise something was amiss. This is a continual process in daily life, but it also applies to a technical environment, where Type 2 thinking could be expected to prevail. Many of the high-level decisions, such as selecting an appropriate modus operandi is done intuitively, using pattern recognition, e.g. building a bridge requires a different mindset to designing a pump, which would be intuitive, although the calculations would involve Type 2 thinking. It would simply be too time consuming (and tedious) to do otherwise. Unfortunately, this also applies to a risk-based modus operandi where the boundaries between science (statistics) and guesswork (prediction) become blurred. In normal life, risk-based decisions are used for anything from crossing the road to making financial investments, so it has become a default mindset that works predominantly with Type 1 thinking. Conversely, it has taken the author around twenty years of Type 2 thinking to identify the eight fundamental errors in prediction, so it is not surprising that practitioners, who have timescales and budgets to adhere to, have not undertaken this work.

Figure 4 Barriers to Change



Human nature involves further barriers to change such as: Confirmation Bias, Overconfidence & Substitution, Theory Induced Blindness, Bounded Rationality, (Kahneman, 2011); Normalisation (Hopkins, 2012), numerical argument preferences (Windschitl, 1996). Confirmation bias is where the individual selects only the evidence that favours their own beliefs. Kahneman demonstrated the remarkable level of confidence that people have in their own risk predictions and also how they tend to substitute a different question to the one they are given when it is not understood or cannot be answered. Theory Induced Blindness is where an initially plausible theory is accepted, but not fully understood. The individual applies the theory, but as understanding improves with time, he or she uses confirmation bias to maintain their 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. Bounded rationality is where that which is not understood is simply excluded from the debate or thinking, e.g. limiting the argument to only those factors that can be quantified. Normalisation is simply the tendency to accept commonplace processes, rather than challenge the ideas. Engineers also have a natural tendency towards numerical quantification as a default means of decision making.

Circular reasoning enables uninformed beliefs to prevail. Because there are no robust statistics for major accidents the models are based on hypothetical algorithms that cannot be verified. They are calibrated by assuming that average plant risks are, say, in the ALARP region. When run on other plants, they give similar results, which are accepted, leading regular users to regard their own judgement as expertise, when it is nothing more than a circular argument.

The most common justification for using QRAs and RAMs is that they are "useful tools", which is of course true if the decision makers choose to believe the numbers, and they rarely have reason to do otherwise. A random number generator might be equally 'useful' if the users are prepared to ignore the source of the numbers.

Pride and fear are also strong disincentives to change. Many people have significant history in prediction and may be reluctant to admit that the errors are much greater their perceptions. Furthermore, producing a WRA may involve unknown factors that could open up a wider debate and expose failings in the case for safety. Fear of failure may be a strong disincentive.

There are also a number of political and technical reasons why the status quo has endured. There is a wealth of guidance for UK regulations such as nuclear, chemical, offshore and rail, which predates the RSS guidelines and much of it is based on R2P2, including its 'carrot diagram'. Risk quantification spreads across many disciplines, so no single engineering institution takes specific responsibility for its governance. Although the RSS guidelines are clear on this subject, they apply to the law courts and do not appear to have been adopted by the engineering institutes. There are conflicts of interest, with a service industry providing risk quantification modelling. Statistical competency is also lacking, as most of this work is done by engineers, who do not have sufficient training in the subject.

The fact that these predictions can never be verified, as they relate to rare events, means that the normal processes in scientific advancement will not apply (Kuhn, 1962). Kuhn showed how paradigm shifts only occur after what he described as a 'crisis', where old paradigms no longer work. The fact that QRA and RAMs are not science is the main reason they have not been challenged. The result is a 'Perfect Storm' of reasons why progress has not been made; why prediction has not been abolished and why WRA is so often not given a high enough profile.

Conclusions

Any demonstration that risks have been reduced to ALARP should be based on an objective, Well-Reasoned Argument (WRA), (or robust statistics for the very rare occasions that they might be available). QRA, PRA and RAMs have been comprehensively debunked, because they contain multiple errors and constitute prediction, which has been shown to be little more than guesswork. Prediction conceals incompetence, by substituting reason with numbers, and it prematurely curtails the analysis of hazards. Any safety insights gained are typically indirect and coincidental and could be better achieved by systematic qualitative approaches. Quantification has been shown to be a potential causal factor in many major accidents. Cost Benefit Analysis cannot be mathematically valid for the errors and uncertainties associated with any unique hazard situation.

Whilst Bow Ties may be effective in communicating barrier concepts and listing safety critical elements, an unambiguous definition of barriers has proven to be elusive. They are too simplistic to represent detailed scenarios, common cause failures, independence and relationships between barriers and this typically leads to an optimistic view of risk.

A legally sound means of demonstrating ALARP will identify areas justifying deeper analysis. This requires a change of mindset, which begins with greater objectivity and clearer criteria for terminating the analysis. It constitutes a WRA, devoid of any prediction, based on foreseeability, societal expectation and reasoned disproportionality, which considers factors such as independence, effectiveness and clustering of barriers, consequence reduction, risk trade-offs, design error and UCAs.

The paper raises the question of whether recent developments in scenario analytical techniques, such as STPA, which has been developed for complex software and aerospace systems, should have more application in other industries, and if so how? Some significant major accidents that were not complex aerospace systems, have been reviewed and found to involve sociotechnical scenarios that may not have been appreciated. Ideas have been posited, which form the basis of more objective and pragmatic methods of incorporating these processes, but these will need testing and further development to prove their practicality.

Despite these lessons, there exists an almost perfect storm of barriers to progress, which have psychological, political and technical roots. As long as the Regulator continues to approve predictive risk assessments, they will remain an attractive means of making a case for safety. Regulatory and industry guidance therefore needs to be urgently overhauled to clearly differentiate between robust statistics and prediction, with prohibition of the latter. This would stimulate significant progress in industrial safety.

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

Baker, J., 2007, The Report of the BP U.S. Refineries Independent Safety Review Panel

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

Buchanan, M., 2001, Ubiquity: Why Catastrophes Happen, ISBN 0-609-80998-9

Cullen, Lord, 2001, The Ladbroke Grove Rail Inquiry, HSE Books, ISBN 0 7176 2056 5

Derksen, T., 2009, 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

Hill, R., 2002, Cot Death or Murder? - Weighing the Probabilities, University of Salford

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

HSE, 2001, Reducing Risks Protecting People, ISBN 0 7176 2151 0

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

Miller, K., 2018, Quantifying Risk and How It All Goes Wrong, IChemE

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