The Science and Superstition of Quantitative Risk Assessment

Andrew Rae[a*], John McDermid[a], Rob Alexander[a]
[a]University of York, York, United Kingdom

Abstract: In safety, environmental, and financial regulation the public are often asked to accept estimates of a concept, "risk", that they cannot directly perceive. Faith in these estimates is supported by logical reasoning but not by empirical evidence. Unfortunately, the evidence that does exist about risk phenomena indicates that human reasoning about risk is highly unreliable.

In this paper we determine what properties must hold for Quantitative/Probabilistic Risk Assessment (QRA) to be fit for purpose. We identify these properties by considering how the outputs of QRA are actually used by engineers and regulators. We then consider what evidence could be realistically available to demonstrate these properties – i.e., to what extent can a particular QRA technique be validated against the properties? We discuss whether it is possible to directly test the properties, or at least to test the arguments made for and against the properties. Against this range of possible evidence, we determine what evidence does in fact exist. We find that whilst it is possible to test whether QRA has the properties expected of it, good evidence is not currently available.

This conclusion should not necessarily be interpreted as evidence against the safety of industries using QRA, but does cast into doubt the extent to which QRA contributes to the achievement of safety. It also suggests that if there are benefits to QRA, there is no evidenced reason to believe that they arise from quantification rather than from the process of systematically analysing the sources of risk.

Keywords: PRA, QRA, System Safety, Research Validation

1 Introduction

1.1 Problem Statement

Engineering is "the application of scientific and mathematical principles to practical ends."[1] Risk Management has a well understood set of practical ends, and a well-populated set of principles and practices for achieving those ends. In this paper we examine the fitness for purpose of the principles and practices related to assigning a numerical quantity to risk. Safety regulation will be used as a running example, but the work is applicable to risk regulation more generally.

The underlying principles of risk assessment are captured in the National Academy of Science "Red Book" [2]. The Red Book models regulatory action as two conceptually distinct activities – assessing risk, and making decisions based on that risk. It treats assessment of risk as an essentially scientific activity, limited by the available scientific knowledge and uncertainty surrounding the risk. It treats decision making in response to the risk as a political activity, with risk assessment as one input.

The Red Book implicitly represents risk as a quantity attached to a particular outcome or set of outcomes. The risk of any particular negative outcome is determined by considering both the severity and the likelihood of the outcome. Risk as a numerical quantity is useful for:

- Making appropriate decisions about risk mitigation measures based on their cost and the extent to which they are likely to reduce risk [3]
- Making trade-off decisions between different sources of risk, such as choosing between energy sources or design options [4]
- Regulating and accepting the risk associated with a particular project [5]
- Transferring risk through insurance or compensation [6]

The term "risk assessment" is used widely in both literature and industry, and covers a range of meanings. In its narrowest sense, a quantitative risk assessment involves determination of the likelihood of a particular event sequence for a given scenario. An example risk assessment of this type is given by Rochelle [7], who calculates the probability of at least one individual being fatally struck by a piece of the Upper Atmosphere Research Satellite (UARS) after it re-enters the atmosphere from orbit. A wider form of risk assessment considers the likelihood of a particular event in all scenarios. An example risk assessment of this type is contained in the "WASH-1400" Nuclear Reactor Safety study [8], which calculates the likelihood of release events of specific types and magnitudes. Broader still are risk assessments which determine the total likelihood of particular consequences arising from a particular decision, such as the Environmental Protection Agency (EPA) determining the risk of fatality if the use of organophosphates is permitted [9]. The widest type of risk assessment attempts to determine the total risk presented by a system or decision (in principle aggregating all possible consequences. Safety regulations for the introduction of equipment in high-risk industries such as railways [10] or aircraft [11] typically require such assessments.

Each increase in the scope of risk assessment adds a layer of uncertainty. This can be readily seen in the case of a deterministic system such as a computer program (considered in the abstract, without reference to hardware). For a particular scenario (a given input sequence), the chance of a particular output can be precisely calculated – in fact, it is either certain or impossible. Determining the overall probability of a particular output requires characterization of the input profile, as well as knowledge of the outputs for each input. Determining the probability of outputs causing a particular consequence adds the requirement to determine which outputs have that consequence. Determining total risk requires additionally knowing all of the real world consequences the software could contribute to.

Quantitative Risk Assessment (QRA) and Probabilistic Risk Assessment (PRA) - hereafter referred to jointly as QRA – are applied at all levels of risk assessment. QRA is widely used, and both explicitly and implicitly required by many standards and regulations. Despite a wide body of research on improving and applying QRA, very little information exists on the validity of the techniques. Without validation, there is limited potential to improve the state of the art or the state of practice, and good reason to view the results provided by QRA with skepticism. In other words, QRA is used to assure the integrity of systems, but QRA has not itself undergone the necessary assurance.

2 Necessary properties of QRA

2.1 Source of Required Properties

There is no academic or industrial source which spells out the properties that QRA must have. However, we can infer the required properties by looking at how QRA is used and what people say about QRA. We state these properties here as claims which can be made about QRA, and which can in theory be supported or refuted by evidence.

PRIMARY CLAIM: That the "top number" – the aggregated measure of total system risk – is sufficiently accurate and precise.

SECONDARY CLAIM: That the process of conducting QRA provides measurable safety benefit compared to equivalent non-quantified risk assessment activities.

FALLBACK CLAIM: That QRA is one of a range of activities that provides measurable safety benefit, but cannot be shown to be better or worse than other activities which cause time to be spent thinking about how to make a system safer.

To check that these claims were a fair representation of the way QRA is used and discussed, we surveyed a collection of real-world risk assessment reports [12] and noted the self-identified purpose of the reports. Activities making use of QRA include:

- Classifying risk (usually for the purpose of regulating a substance or technology)
- Reacting to public concern regarding a known or suggested risk
- Identifying ways to improve a design
- Selecting between competing designs

- Comparison of risk with pre-defined benchmarks
- Trading-off risk against other concerns
- Operational risk management
- Accepting or declining risk as a public policy decision

2.2 How Precise and Accurate Does the Top Number Need to Be?

The WASH-1400 report [8] and its successor NUREG 1150 [13] formed the basis of nuclear regulation in the United States over the past decades [5]. WASH-1400 assigned frequencies to various radiation release events ranging from 9×10^{-7} to 1×10^{-4} per reactor per year. Uncertainty ranges for these figures were not provided, but there is an implicit claim in stating the 9×10^{-7} figure that the probabilistic methods used were capable of distinguishing between 8×10^{-7} and 1×10^{-6} releases per reactor per year. WASH-1400 is widely argued (see for example Leveson and Rahn [14]) to make pessimistic estimates. If future estimates are to be less pessimistic, they will also need to be more accurate.

It is not just probabilities that need to be accurate. Environmental risk quantification began and continues to be dominated by the adverse health consequences of exposure to toxic or pathogenic agents. Typically this involves a single chemical with specified health endpoints. For example, the European Union risk assessment of "4-Nonylphenol (branched) and Nonylphenol" [15] compares expected population doses to the LD50 (the dose with a 50% likelihood of causing death in an individual). The methods used need to identify the population doses with sufficient accuracy to judge whether the maximum expected dose approaches the LD50.

In the case of QRA used to determine regulatory compliance, the total error must be less than the margin between the estimated risk and the risk limit. For example, if an aircraft system must exhibit a dangerous fault less than 1×10^{-6} per flight hour, and is estimated to exhibit such faults at a rate of 8×10^{-7} per flight hour, then the error must be less than 2×10^{-7} per flight hour. Otherwise the QRA could appear to show that an unsatisfactory system met the target.

2.3 Can We Tell if QRA has been Properly Conducted?

Where QRA is used for making decisions, there must be some way to tell that it has been properly performed. The U.S. Nuclear Regulatory Commission has published guidance on determining whether any particular QRA is technically adequate [16]. This guidance relies on the adequacy of "good industry practice" and the ability of review staff to identify omissions or inappropriate treatment of uncertainty.
Such judgment is difficult, however, because risk assessment is prone to subtle error. The textbook "Misconceptions of Risk" by Terje Aven [17] details nineteen fundamental misunderstandings capable of undermining an assessment. Assuming that Aven is not attacking straw-men (the authors of this paper have personally witnessed many of these misunderstandings in industrial practice, as has Hansson [18]) this places a significant burden on review processes.

2.4 Can QRA Results Survive Communication?

Alexander and Kelly [19] highlight the importance of probabilities and other numbers in communicating risk. Expressing risks numerically is necessary for prioritization, proportional response and aggregation of risk. Whilst Alexander and Kelly argue predominately in favor of QRA, they note that where stakeholders are likely to misinterpret or misuse quantitative estimates, provision of such estimates can be dangerous. Adapting the approach of Watson [20], they suggest that the usable product of QRA is an expressed relationship (an "argument") between input and output estimates. The audience of QRA is expected to review, challenge and/or use the argument to manage safety, rather than merely accept the output estimates.

Whilst this approach is internally consistent, and overcomes many of the epistemological problems with risk estimates and QRA methods, it introduces the requirement that those who conduct QRA communicate the outputs in this fashion, and that those who receive QRA outputs are able to interpret them appropriately.

This expectation is subtle, and not necessarily practicable. A risk practitioner might say "the risk of an accident is 1 x 10^{-8} per year if the estimated failure rates hold and the model is correct", meaning that significant work is required to monitor the failure rates and validate the model. Their audience might simply hear "the risk of an accident is 1 x 10^{-8} per year".

2.5 Does QRA Lead to Safer Systems Even When the Top Number is Not Accurate?

In 2004, Apostolakis presented an enthusiastic defense of QRA in response to popular press skepticism [21]. Notably, he chose specifically not to defend the quantification of aggregate risk, instead pointing to the improvements in safety that may arise from performing the analysis. This is not to say that safety improvement is a side-effect of quantification – Apostolakis suggests that a focus on quantification was necessary:

- to appropriately prioritise risk from different scenarios;
- to facilitate communication between stakeholders;
- to integrate information from different academic disciplines;
- to highlight areas of uncertainty where more information or research is necessary.

Apostolakis also argues for the central role of quality review in achieving these goals. The argument made by Apostolakis is similar to that of Alexander and Kelly [19]. Whilst Alexander and Kelly focus on communication of a faithful mental model of the risk of a complete system, Apostolakis concentrates on identifying opportunities for design improvement or further assurance.

3 Direct Evidence for and against QRA Fitness-for-Purpose

The authors have carried out an organized search for literature empirically addressing whether PRA techniques have the required properties. This search was conducted in two stages. First, an automated search paired keywords representing QRA or specific QRA techniques, with keywords relating to the properties in Section 2 or to evidence gathering and evaluation. Second, a manual search was conducted using queries on mailing lists and tracing references from papers that advocated QRA or specific QRA techniques.

As suggested by our previous work on system safety research validation [22], there is a very low volume of research providing direct evidence either for or against QRA satisfying the fitness-for-purpose requirements. In fact, the only work directly on topic discusses the lack of evidence rather than providing evidence in either direction. Manion [23] discusses the evidence basis for Fault Tree Analysis (FTA), but the only direct statement about FTA performance, that fault trees underestimate risk "often by many orders of magnitude" is in fact a hypothetical assertion from Mauri [24] rather than an evidenced claim. Danielsson [25] discusses quantitative models for financial risk assessment, demonstrating that the paucity of evidence for QRA is not confined to system safety.

The lack of evidence supporting QRA is not due to fundamental difficulties with evaluating QRA performance. Whilst it remains challenging to conduct a controlled experiment to verify whether predicted risk reliably matches actual system risk, there are many relevant questions that *can* be addressed through experimental or field research.

A sample of questions which could be answered through controlled experiment are:

- How good are typical reviewers at spotting known errors within real QRA?
- To what extent is the result of QRA for a given system variable according to the particular practitioner or organisation who performs the analysis?
- To what extent do external influences such as a known risk target influence the conduct and outcome of QRA for a given system?
- To what extent can reviewers identify which QRA is more correct when two QRAs are performed for the same analysis? (Or, failing that, to what extent can reviewers agree which QRA is more correct?)
- Does the representation of QRA influence the understanding of a reviewer or customer?
- Do the quantification steps in QRA improve understanding of risk over and above the non-quantification steps?

A sample of questions which could be answered through field observation are:

- How often do known types of errors occur within real QRA?
- How often does design-time QRA omit or understate hazards which later cause accidents?
- To what extent is QRA used as an in-service risk model?

4 Indirect Evidence for Foundations and Criticism of QRA

Objections to the use of QRA fall into three broad camps:

1. Disagreement that it is possible (or even necessarily desirable) to separate the risk assessment and decision making processes [26];
2. technical concerns about the accuracy and precision of the risk assessment [23]; and
3. concerns about the communication and use of risk assessment results, particularly with respect to uncertainty in the results [27].

These concerns may be addressed by turning them into questions which can be resolved through empirical enquiry.

4.1 Can risk assessment be practically separated from risk decision making?

The separation of "scientific" risk assessment from "political" risk decision making, as advocated by the US National Science Council [2] relies on the ability to procedurally and intellectually separate the two activities. Whilst there is strong evidence that the separation is not maintained in practice (see for example [12] and [26]) this does not directly show that separation is not practicable. Of greater concern is the large body of work relating to the unreliability of human judgment. Of particular concern are *framing effects* and *external signal effects.*

Framing effects relate to distortions in subjective judgment caused by the manner in which the problem is presented. Specific examples observed in risk models are pruning bias and partition dependence. Pruning bias, first described by Fischoff in the context of fault trees [28], causes collected risks to be estimated lower than the same risks considered individually. In the more general case of partition dependence [29], study participants tend to treat all risks as equal unless they have specific information. For example, if asked to compare risk A with risk B, they will tend to evaluate the risks as equal. However, if risk B is instead presented as four separate risks (C, D, E, and F), they will treat A, C, D, E and F as equal, effectively rating risk B as four times more significant than risk A.

An external signal effect is where a decision maker receives a suggestion as to the "right answer" from some source outside the immediate problem. A typical example is when the expected answer to a survey question is suggested by other questions in the same survey [30]. External signals have a practical impact on decision making when regulatory procedures and targets influence expert judgment [31]. If a particular likelihood or severity associated with an adverse event would cause that event to be classified as "unacceptable risk", this is fact should not in theory change the estimate of likelihood or severity, but in practice will do so.

4.2 Can causal models sufficiently represent the pathways to accidents?

There is a distinction between what can go wrong with causal models (causing the model to inaccurately represent the possible paths to accidents), what has gone wrong in the past, and what will inevitably go wrong. We have found no source that disputes the many possible errors in QRA – writers who support QRA implicitly or explicitly judge that these errors can be avoided. In the following analysis we consider cases where there is academic concern about a particular type of error (something that can go wrong) and practical evidence that it can occur in real analysis (it has gone wrong). We consider that in these cases there is a burden of proof requiring evidence that the error can be systematically avoided – in other words, the null hypothesis is the potential for error is sufficient to invalidate QRA.

Problem 1: Incomplete identification of the range of dangerous outcomes
Hazard identification is not a deterministic activity, and there is no possible test for completeness of any hazard list. Industry specific methods of "automated hazard identification" (see Catino [32] as a representative example of a large body of work) rely on matching system behaviors against previously-

identified hazards. Measuring performance of hazard identification as a research exercise is complicated by instances of the same hazard represented in different ways, and the lack of a reference list against which to compare identified hazards. Proving a hazard list to be incomplete is much easier, requiring only examples of hazards which are real but have not been identified. Carter and Smith [33] have performed this exercise for construction industry hazard lists, and Suokas and Pyy [34] conducted a similar investigation of chemical process plants. Both studies found significant gaps between the list of hazards which the authors believed should reasonably have been identified, and the actual lists of hazards. Suokas and Pyy drew the further conclusion that hazard identification was strongest in the case of physical failures, and weakest in the case of management and design problems.

Problem 2: Invalid Assumptions of Independence

Performing QRA calculations requires either statistically independent basic events (e.g. Fault Tree Analysis) or known conditional probabilities (e.g. Bayesian Belief Networks). If the technique assumes independence but this assumption is not true, top event probabilities are not valid. Violations of independence can be viewed more intuitively as "common-cause failures" (CCFs) – a single unidentified basic event which couples two identified events.

Beer [35] reviewed 609 aircraft accident and incident reports and found that common-cause failures were present in 11% of the reports. The 11% included only cases where Beer could positively confirm the existence of CCF. Unfortunately accident reports seldom refer back to the original risk assessment, so this does not demonstrate that the original QRA omitted the CCF. However, the fact that an accident occurred is strong circumstantial evidence that the CCF escaped appropriate treatment.

Problem 3: Unanticipated System Behavior

Joyce and Wong [36] describe the problem of system behaviors which are "not non-conformant" – i.e., they do not strictly contradict any identified requirement, but are still undesirable. They argue that unsafe, not non-conformant behaviors can occur in the presence of structured safety programs, and therefore that testing against safety requirements is insufficient for safety assurance.

Leveson [37] describes the Arianne 501 launch failure, the loss of the Mars Polar Lander and a UK chemical plant accident as industrial instances where all individual components behaved exactly as specified, leading to a dangerous system state.

A specific cause of unanticipated system behavior is where a system behaves "as-designed" but in an unanticipated environment. The Fukushim loss of cooling accident [38] is an example of credible environmental circumstances beyond the design-basis used for QRA which never-the-less led to an accident.

Problem 4: Omission of a Category of Concerns

Anecdotally there is a high incidence of QRA omitting entire categories of concerns such as software, human error, maintenance effects or environmental factors. In such cases the QRA may accurately represent the likelihood of accidents from some causes, but inaccurately communicate it as the likelihood from all causes.

Every accident or incident involving a system previously thought to be safe provides circumstantial evidence that the underlying failure model was incomplete, although it is not always possible to differentiate between technical and managerial error in this regard. In as yet unpublished work, Reinhardt [39] describes two Royal Air Force investigations into unreliable systems which found failure modes not considered in the original quantitative analysis.

4.3 Can inputs to QRA be accurately quantified?

Skeptical critiques of quantitative risk assessment (see Manion [23], Hansson [18], Apostolakis [21], Peter [27]) emphasize the poor pedigree of the source data. In the above discussion we hope to make clear that source data uncertainty is only one of several challenges to the utility of QRA. However, this does not take away from the legitimate concerns about "Garbage In, Garbage Out". Unless divergent data sources are used

to triangulate probabilities (a rare practice in QRA) combination of uncertain sources can only increase, rather than decrease, the associated uncertainty. The common use of single-point data instead of more appropriate probability distributions compounds the use of poor source data.

Specific concerns include:

- Incorrect anticipation of the range or frequency of human error [40]
- Incorrect modeling of complex processes such as evacuation, fire, flooding or structural movement [41]
- Unreliable expert judgment [29][42]
- Insufficient historical data [43]
- Over-optimistic assessment of mitigations [44]

The Lewis Committee Report [45] produced in response to criticism of WASH-1400 raised concern not only that the uncertainty of the WASH-1400 conclusions was unstated, but that it was unknown. Lewis did not reject the use of probabilistic risk assessment, although one of the authors of the report questioned whether the methods could ever provide estimates with sufficient confidence. Of particular concern to the Lewis Committee was inability to properly quantify human error or common cause failures.

4.4 Do QRA results maintain validity beyond system deployment?

QRA is a "snapshot" of system risk at a particular time. If the system risk varies significantly over time, QRA can be very misleading about the actual risk of a system. Factors which threaten the validity of QRA as a system is used are detailed in Rae and Nicholson [43]. Of particular concern are:

- Divergence between the modeled system and the as-built system;
- Interactions between social and technical parts of the system which invalidate technical assumptions [46]; and
- Changes in effectiveness of mitigations over time.

5 Conclusion

5.1 What can be reasonably concluded from the evidence?

In this paper we have shown that there are a set of properties which QRA must have to be fit-for-purpose. In establishing these properties we have taken into account the multiple purposes that QRA has, with the understanding that it may be fit for some purposes but not for others. These properties must hold for QRA *as it is actually practiced*. It is insufficient for QRA to be merely *capable* of being fit for purpose, it must actually be so.

Our contention that empirical evidence is necessary to establish the properties of QRA is based on two lines of argument:

- There is no extant evidence that QRA has the required properties, despite opportunity for such evidence to be produced
- There is a body of evidence to support the existence of multiple problems with QRA, placing a positive burden of proof on proponents and users of QRA to show that these problems can and have been dealt with

In the absence of systematic evidence for or against QRA efficacy, we can classify QRA properties as follows:

With respect to:

PRIMARY CLAIM: That the "top number" – the aggregated measure of total system risk – is sufficiently accurate and precise.

This claim is not evidenced and is implausible in the general case of total risk. It is plausible that it may be true for specific outcomes arising through specific mechanisms, as many of the challenges to QRA do not apply in these cases.

With respect to:

SECONDARY CLAIM: That the process of conducting QRA provides measurable safety benefit compared to equivalent non-quantified risk assessment activities.

This claim is poorly evidenced, but plausible mechanisms are proposed by Aven [17] and Apostolakis [21].

With respect to:
FALLBACK CLAIM: That QRA is one of a range of activities that provides measurable safety benefit, but cannot be shown to be better or worse than other activities which cause time to be spent thinking about how to make a system safer.

This claim lacks systematic evidence, but is highly plausible and is not challenged by any of the indirect evidence against QRA efficacy. In the absence of systematic evidence, the accumulated anecdotal experience of safety practitioners lends weight to the claim.

These conclusions concern the evidence for QRA, which is itself used as an evidence-generating activity. The fact that QRA lacks evidence does not mean that outcomes of QRA are incorrect, merely that there is insufficient reason to trust them. It is not evidence that the real risks are higher or lower than QRA estimates, but it is reason to be concerned about the accuracy of those estimates.

5.2 Where to from here?

Particular decisions made with the assistance of QRA are 2nd order removed from the evidence deficit. Problems with QRA are not problems with the industries which tend to use QRA. An industry which tries hard to improve safety using an imperfect method may still be safer than one which doesn't try as hard, simply by virtue of the attention paid to safety, and the consequential design improvements. For industries such as aviation and nuclear power, there is direct evidence from accident rates to shows those industries in good light independently of their use of QRA.

Our concern is that the low accident rates in these industries may be independent of, or only indirectly linked to, the use of QRA. This is important for other industries attempting to emulate the success of nuclear power or aviation. Promotion of system safety in these other industries should focus on those practices which are most likely to add real value, and for industries such as health care may require strong empirical evidence to gain traction.

Knowledge about the efficacy of QRA is important also for efficient allocation of effort. With trained and experienced safety practitioners in high demand, their time may be better spent on activities other than QRA. The most obvious alternative is spending time qualitatively understanding the hazardous behaviors of the system.

Our investigation supports the following actions:

Firstly, it is evident that numerical claims supported by QRA must be moderated. In particular, when QRA users communicate the results of QRA, they need to more strongly express the uncertainty inherent in the methods themselves, not just the uncertainty propagated from the inputs to the outputs.

Secondly, there is a need for research to fill the evidence gaps associated with QRA. It is likely that positive direct evidence of QRA inaccuracy will be necessary to halt current misuse of QRA. Additionally, the hypothesis that QRA has intrinsic benefits independent from the numerical outputs should be tested in both controlled experiments and practical settings.

Finally, the QRA community needs to make sure that industries undergoing safety reform, such as deep-sea oil and gas extraction, learn the right lessons from successful industries. This will require rigorous evaluation of QRA practices against their explicitly claimed benefits.

References

[1] *The American Heritage Dictionary of the English Language*, Fourth ed. Houghton Mifflin, 2000.

[2] Committee on the Institutional Means for Assessment of Risks to Public Health, National Research Council, *Risk Assessment in the Federal Government: Managing the Process*. Washington, D.C.: The National Academies Press, 1983.

[3] Health and Safety Executive, *Reducing Risk Protecting People*. HSEBooks, 2001.

[4] C. A. Cornell, 'Bayesian Statistical Decision Theory and Reliability-Based Design', *Structural Safety and Reliability*, no. 254, 1972.

[5] N. Siu and Dorothy Collins, 'PRA Research and the Development of Risk Informed Regulation at the U.S. Nuclear Regulatory Commision', *Nuclear Energy and Technology*, vol. 40, no. 5, pp. 349–364, Aug. 2008.

[6] G. O. Robinson, 'Probabilistic Causation and Compensation for Tortious Risk', *J. Legal Stud.*, vol. 14, p. 779, 1985.

[7] W. Rochelle, J. Marichalar, and N. Johnson, 'Analysis of reentry survivability of UARS spacecraft', *Advances in Space Research*, vol. 34, no. 5, pp. 1049–1054, 2004.

[8] United States Nuclear Regulatory Commission, 'Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants'. Oct-1975.

[9] Environmental Protection Agency, 'Organophosphorous Cumulative Risk Assessment'. 31-Jul-2006.

[10] G. M. H. Department for Transport, 'European Railway Safety Directive', 23-Oct-2007. [Online]. Available: http://www.dft.gov.uk/pgr/rail/Safety/ersd. [Accessed: 06-Apr-2011].

[11] ISO - International Organization for Standardization, 'Safety Aspects: Guidelines for their Inclusion in Standards', ISO/IEC Guide 51, 1999.

[12] A. Rae and R. Hawkins, 'Risk Assessment in the Wild', presented at the Australian Safety Critical Systems (submitted), Brisbane, 2012.

[13] Office of Nuclear Regulatory Research, 'Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants - Final Summary Report', United States Nuclear Regulatory Commission, NUREG-1150, Dec. 1990.

[14] M Levenson and F Rahn, 'Realistic Risk Estimates', *IAEA Bulletin*, vol. 23, no. 4, 1981.

[15] European Chemicals Bureau, 'Risk Assessment Report - 4-nonylphenol (branched) and nonylphenol'. European Communities, 2002.

[16] Office of Nuclear Reactor Regulation, 'Determining the Technical Adequacy of Probabilistic Risk Assessment Results for Risk-Informed Activities (NUREG-0800, Chapter 19.1)'. U.S. Nuclear Regulatory Commission, Nov-2002.

[17] T. Aven, *Misconceptions of Risk*. Wiley-Blackwell, 2010.

[18] S. O. Hansson, 'Seven Myths of Risk', *Risk Management*, vol. 7, no. 2, pp. 7–17, Jan. 2005.

[19] R. D. Alexander and T. P. Kelly, 'Escaping the Non-Quantitative Trap', in *Proceedings of the 27th International System Safety Conference*, 2009.

[20] S. R. Watson, 'The meaning of probability in probabilistic safety analysis', *Reliability Engineering & System Safety*, vol. 45, no. 3, pp. 261–269, 1994.

[21] G. E. Apostolakis, 'How Useful Is Quantitative Risk Assessment?', *Risk Analysis*, vol. 24, no. 3, pp. 515–520, Jun. 2004.

[22] A. J. Rae, M. Nicholson, and R. . Alexander, 'The State of Practice in System Safety Research Evaluation', presented at the IET System Safety, Manchester, 2010.

[23] M. Manion, 'The epistemology of fault tree analysis: an ethical critique', *International Journal of Risk Assessment and Management*, vol. 7, no. 3, 2007.

[24] G. Mauri, 'Integrating Safety Analysis Techniques, Supporting Identification of Common Cause Failures', PhD, University of York, York, 2000.

[25] Jon Danielsson, 'Blame the Models', *Journal of Financial Stablity*, Jun. 2008.

[26] E. K. Silbergeld, 'Risk assessment: the perspective and experience of U.S. environmentalists', *Environmental Health Perspectives*, vol. 101, pp. 100–104, Jun. 1993.

[27] M. Peter, 'Reducing the harms associated with risk assessments', *Environmental Impact Assessment Review*, vol. 24, no. 7–8, pp. 733–748, Oct. 2004.

[28] B. Fischhoff, P. Slovic, and S. Lichtenstein, *Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation*. Defense Technical Information Center, 1977.
[29] C. R. Fox and R. T. Clemen, 'Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias toward the Ignorance Prior', *Management Science*, vol. 51, no. 9, pp. 1417–1432, 2005.
[30] J. A. List, 'Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards', *The American Economic Review*, vol. 91, no. 5, pp. 1498–1507, Dec. 2001.
[31] K. Bawn, 'Political Control Versus Expertise: Congressional Choices about Administrative Procedures', *The American Political Science Review*, vol. 89, no. 1, pp. 62–73, Mar. 1995.
[32] C. A. Catino and L. H. Ungar, 'Model-based approach to automated hazard identification of chemical plants', *AIChE Journal*, vol. 41, no. 1, pp. 97–109, Jan. 1995.
[33] G. Carter and S. D. Smith, 'Safety Hazard Identification on Construction Projects', *Journal of Construction Engineering and Management*, vol. 132, no. 2, pp. 197–205, Feb. 2006.
[34] J. Suokas and P. Pyy, 'Evaluation of the validity of four hazard identification methods with event descriptions', 1988.
[35] J. A. Beer, 'The True Significance of Common Cause Failures in Accidents', Masters Thesis, University of York, United Kingdom, 2011.
[36] J. Joyce and K. Wong, 'Hazard-Driven Testing of Safety-Related Software', presented at the 21st International System Safety Conference, Ottawa, 2003.
[37] Leveson, 'A new accident model for engineering safer systems', *Safety Science*, vol. 42, no. 4, pp. 237–270, 2004.
[38] C. Perrow, 'Fukushima, risk, and probability: Expect the unexpected', *Bulletin of the Atomic Scientists*, 01-Apr-2011.
[39] D. Reinhardt, 'TBD', Draft PhD Thesis, University of York, 2012.
[40] J. Rasmussen, 'Risk Management in a Dynamic Society: a Modelling Problem', *Safety Science*, vol. 27, no. 2–3, pp. 183–213, 1997.
[41] Hanea, Ale, Jagtman, Safety Science Group, and TU Delft, 'Human Risk of Fire: Building a decision support tool using Bayesian networks'. 27-Nov-2009.
[42] A. Mosleh, V. M. Bier, and G. Apostolakis, 'A critique of current practice for the use of expert opinions in probabilistic risk assessment', *Reliability Engineering & System Safety*, vol. 20, no. 1, pp. 63–85, 1988.
[43] A. J. Rae and M. Nicholson, 'Making robust ALARP decisions for in-service systems', 2009, p. 1A2–1A2.
[44] M. D. Watkins and M. Bazerman, 'Predictable Surprises: The Disasters You Should Have Seen Coming', *Harvard Business Review*, Mar. 2003.
[45] H. W. Lewis, R. J. Budnitz, W. D. Rowe, H. J. C. Kouts, F. von Hippel, W. B. Loewenstein, and F. Zachariasen, 'Risk Assessment Review Group Report to the U. S. Nuclear Regulatory Commission', *IEEE Trans. Nucl. Sci.*, vol. 26, no. 5, pp. 4686–4690, 1979.
[46] C. Perrow, *Normal Accidents: Living with High-Risk Technologies*, Updated. Princeton University Press, 1999.