Simulation and Prediction in Safety Case Evidence

R. D. Alexander, PhD, T. P. Kelly, DPhil; University of York; York, England

Keywords: safety analysis, safety management, certification, epistemology

Abstract

The use of simulation in safety analysis is often criticised. We are faced, however, with safety critical systems of ever-greater complexity, and with the demand to extend safety engineering to cover large-scale sociotechnical systems such as hospitals and military forces. Simulation methods offer ways to analyse aspects of these systems that other, ‘traditional’ methods cannot, because of size, complexity or non-linear behaviour.

It would be foolish to claim that simulation techniques aren’t frequently misused, or that there haven’t been serious errors influenced by poor simulation. Every method of analysis and prediction, however, has its own risks, flaws and limitations. The solution is twofold; first, an engineer must know the specific capabilities and limits of the methods that they use, and hence know what information they can give and how they complement each other. Second, an operator must update their predictions (from both accidents and near-miss incidents) as part of their safety management system, thereby evaluating both the correctness of their predictions and the effectiveness of their tools.

In this paper, we review the general strengths and weaknesses of simulation methods, and compare these to the capabilities of other several other methods. We raise the common vulnerability of all the methods to changes in the world, and draw some conclusions for safety analysis in general.

Introduction

There are number of existing papers (such as (ref. 1, 2, 3)) that promote the use of simulation in system safety engineering, particularly for safety analysis. Typically, these papers make one of three claims: that simulation allows a greater number of situations to be considered in analysis, that it allows an improved understanding of *why* a system behaves in a certain dangerous way, or that it allows analysis where none was possible before. Some authors have criticised simulation, however, for analysis and prediction in general (ref. 4), or for safety analysis in particular (ref. 5).

When an engineer is building a safety case for a system, the claims in their argument need to be supported by evidence, and simulation is one of many sources for such evidence. Other sources of evidence include expert opinion, statistical models, and experiments or field trials using the actual system. As with simulation, these have their own pattern of strengths and weaknesses. The assurance that an identified item of evidence (such as the results from a software test suite, or the output of a mathematical model) contributes to a safety case depends on the specific claims that it is used to support (such as the dangerous failure rate under certain operational circumstances). Importantly, this is not a simple function of ‘level of assurance’ needed – it is the *nature* and *kind* of claim that matters, the fit between the *type* of evidence and the *type* of claim.

It follows that the use of simulation evidence in a safety case is appropriate, and indeed desirable, when the claim being made is within the capabilities of simulation evidence. It is particularly important when other sources of evidence cannot provide adequate coverage of the space of concern. This paper reviews the general strengths and weaknesses of simulation methods, and shows how these relate to the capabilities of several other methods.

The next section introduces the idea of prediction as the central core of safety analysis (and, by extension, safety cases). It is then shown how simulation can give insight where other methods cannot, while at the same time being open to a number of dangerous misuses. This is followed by a review of a range of alternative evidence sources – the strengths and weaknesses of these are discussed, and their relationship to simulation is made clear.

The Need for Prediction

At the core of a claim about the safety of a system, there is always an appeal to prediction of the future. When a safety engineer attempts to certify a system for operation, he is typically meeting a standard that defines what is necessary to claim that a system is safe. When an explicit safety case is prepared (as required, for example, by the UK military safety standard Def Stan 00-56 (ref. 6)), the ultimate top-level claim is, as a rule, a claim that the system is acceptably safe given its expected use. In order to be of interest to the assessor of a safety case (who will approve the system for certification on the basis of that case), this claim must be interpreted as one that the system will prove to be safe *in practice*, during an operational lifetime that extends years, possibly decades into the future. This is a prediction.

The detail is a little more complex than that – a safety case will state (or, in a substandard safety case, just leave unstated) certain assumptions about the operating context of the system. At levels below the main claim, there will be claims such as “If hazard X occurs, then accident Y will occur with probability 10^{-2}”. These are predictions under certain limited contexts, but predictions nonetheless (their presence is valuable because they are easier to challenge than the top-level claim). For adequate safety, however, they must nevertheless be summed to give an overall assessment of the safety of the system – see Arntsen in (ref. 7). The summation of risks over all hazards, for example, is part of MIL-STD-882E (ref. 8).

It is important to note that we are not referring here to *point predictions* – a safety case need not make a claim that “there will be 17 non-trivial accidents during the lifetime of the vehicle fleet, resulting in 5 deaths and 40 serious injuries”. That’s impracticable and unnecessary. Rather, we are concerned with predictions that the behaviour of the system will display a certain general trend; that certain measureable properties will fall into certain ranges. Typically, this takes a form akin to “fewer than 10^{-6} fatal accidents per operating hour”. Once we have asserted this at the top of our safety case, we then develop the argument that we can *predict that this is going to be true in practice*.

It can even be argued that every decision implies a prediction. For example, if we make a decision to certify a new aircraft engine, then we are predicting that it will prove adequately safe in the future. At a lower level, as the designers of that engine, if we decide to use a particular new type of fan blade on the grounds that “it will be safer” then we are making the prediction that an engine using that type of blade will exhibit fewer accidents, in reality, than one that uses an alternative type[1]. A decision implies the claim that something of value to the decision maker will be better in the future given the option that is chosen, as opposed to the alternatives.

Simulation in Safety Analysis

Simulation techniques are valuable in that it can provide a way to derive the behaviour of models that have highly nonlinear behaviour, that have a huge space of possible parameters (whether these are configuration settings or circumstances of use) or that simply have complex internal dynamics. It is particularly useful for complex systems where there is a high level of ‘contingency’ or path-dependency – events in the system at time t have a dramatic effect on the behaviour of the system at time $t+100$ (see Law in (ref. 9) and Ilachinski in (ref. 10))

The class of simulations known as multi-agent models are particularly powerful – they can combine a variety of disparate representations of system entity behaviour (including equations and decision rules) in order to describe systems where the overall system behaviour emerges from the interaction of highly autonomous entities (most large sociotechnical systems are of this type). Effects arising from ad hoc interactions between entities, which are highly contingent on the state of individual entities and on the timing of the interaction, can be modelled.

The above is particularly relevant given the increasing use, and interest in, complex large-scale sociotechnical systems such as decentralised air-traffic control and network-enabled military units.

[1] There is a second component to such decision-making: one must propose possible options before they can be evaluated.

Simulation can provide a surrogate for experiments when actual experiments are difficult or impossible to perform. For example Johnson, in (ref. 2), talks about the use of multi-agent simulation for modelling building evacuations after fires or bomb threats. For large buildings, especially where the public is involved, performing evacuation drills is dangerous. In the case of hospitals, the risk to patients is so great that adequate evacuation drills are rarely performed. Johnson's simulations allow an attempt to predict what would happen if such a drill *was* performed. Such simulations may be particularly valuable in giving decision-makers cause for concern – they may provide suggestive evidence that the system is unacceptably dangerous. To express this in another way: an un-evaluated simulation is an undesirable thing to have, but a real-world system where we have no clue as to certain critical behaviours is worse.

The critical role that simulations (and similar models) play is to allow the exploration and analysis of system dynamics that are complex and intricate. It can also be observed that we, as a civilisation, have greater computing power than ever before available, and this allows simulation to be used in ways that were not possible in the past. There is, therefore, potential for simulation to allow us to build safe systems with more complex dynamics than was possible in the past.

There are, however, a great many objections to the use of simulation in any important capacity. One of the most commonly raised objections is that the behaviour of a simulation depends on a huge number of assumptions that are both difficult to validate and, in any case, invisible to users of the simulation model who were not involved in its development. A second objection is so-called 'epistemic opacity' (a term used by Humphreys in (ref. 11)) – when a simulation produces a result, it is often difficult to know how that result arose from the internal mechanism of the simulation engine and model. A third, and perhaps most pernicious, objection is that it is easy for a simulation model to be overly convincing to non-specialists, and that this credibility can be independent of actual validity. Roman, in (ref. 12), refers to the case of a low-validity simulation given an attractive animation as GIHO: *"Garbage in, Hollywood out"*.

Multi-agent simulations are particularly opaque, mainly due to the high degree of 'contingency' they exhibit – an apparent insignificant change to a parameter (or to a random number generated early in a simulation run) can have a major effect on the outcome. This leads to an output landscape that is hard to explore, raising doubts about the completeness of any analysis (this is with respect to the model itself, before the correspondence of the model to reality has even been considered). This is discussed further by Richardson in (ref. 13).

Poor use of simulation has been implicated in several significant failures. A pertinent example of incorrect assumptions is the modelling of impact on the space shuttle performed prior to the Columbia accident – an assumption was made that any impacting foam fragment would strike the main wing panels, but in the event the impact was against the leading edge of the wing (ref. 14). Several other examples of failures by computer models in general (not just simulation) can be found in Oreskes and Belitz (ref. 15).

Models Can *Always* be Wrong

Simulation aside, there are a great many other ways to analyse systems and to predict their safety-critical behaviour. In a safety case, these are potential alternative sources of evidence. In this section, we will discuss several such sources, and outline their strengths and weaknesses with an emphasis on how they relate to simulation.

Mental Models

Much of engineering, and indeed everyday life, relies on inferences drawn from models that exist only informally, in the minds of their users. This is the realm of "engineering judgment", based on past experience with similar systems, intuitions about how the world works, and general principles gained through socialization in the engineering culture. There is a large amount of published literature and rhetoric emphasizing the importance of these for safe, effective engineering (a typical example being Ferguson in (ref. 5)).

Partly, the motive for this literature is political, stemming from the role of "engineering judgement" in establishing the credibility and importance of experienced engineers. There is however, strong evidence that engineers can often foresee the consequences of design decisions. In the safety field, this is exemplified by the strong track record of hazard analysis techniques such as HAZOP (ref. 16), where teams of engineers manage to predict a large proportion of the accidents (and causes of accidents) that will be experienced during the operation of a complex engineered system (for HAZOP, typically a chemical process plant).

There is also a large literature, however, on the limitations of human intuition and informal decision-making. For example, intuitive reasoning is known to be biased towards confirming already-held theories, assigning implicit probabilities to events based on their vividness to the imagination, and bizarre phenomena such as ‘anchoring’ (where a person’s answer to a numerical question is influenced by a number that they recently encountered in a wholly unrelated context). A good survey and starting point is the book by Hastie and Dawes (ref. 17).

There is now evidence from many domains that human experts often fail to live up to their own impression of their abilities, particularly when it comes to prediction. The first significant study in this vein was Meehl (ref. 18) in psychiatry, but there is a tradition of similar studies (see Grove and Meehl in (ref. 19) for a meta-analysis of such studies in clinical psychology). The authors are not aware of any such studies in the field of engineering specifically, but the overriding impression from other domains is one of tremendous overconfidence by experts in their intuitive abilities. (There are some controversies and caveats here. For example, there appears to be a strong connection between task characteristics and the effectiveness of expert judgement – see Shanteau in (ref. 20)).

It can be observed that the changes in the systems to which system safety engineering is being applied (as discussed earlier in the paper) are problematic for the use of intuitive judgement. Ever-larger, ever-more-complex systems, which are intentionally ever-more decentralised. In particular, decentralised systems that function via complex interactions between their parts seem to be inherently difficult for humans to understand – Resnick gives some examples in (ref. 21). We already know that HAZOP analyses (and similar manual hazard analysis techniques) do not indentify all hazards in process-plant systems (ref. 16). It seems likely that they will do worse in complex, decentralised systems – Shanteau, in (ref. 20), makes a rudimentary attempt to classify tasks in terms of how well human experts will perform them (when compared to non-experts), and tasks involving non-linear phenomena are problematic.

Statistical Models

Statistical models can be used to extrapolate future behaviour from past behaviour. The most common example is extrapolating the future accident rate of a system from the accident rate observed in its operation so far. More contentious, though potentially more useful, are attempts to extrapolate from the behaviour of one system to the behaviour of a new system (for which actual accident data is of course not available). At a more general level, attempts have been made to justify systems as safe based on statistical evidence of error rates previously observed given the development methods used.

The key advantage of statistical models over intuitive judgement is that the numerical data used and the statistical models that are derived from it are explicit and can be reviewed, analysed, and exchanged without loss between users. In the judgement and decision-making research discussed above, mathematical models are the most common alternative proposed – their use in prediction is commonly referred to as the ‘actuarial method’ (see Hastie and Dawes in (ref. 17)). They have proved better than informal expert judgement in a great many cases (e.g. see Grove and Meehl in (ref. 19)).

Statistical inferences can, of course, be incorrect, and statistical models can be unsound. There are a wide variety of errors that can be made in building mathematical models. As with simulations, assumptions need to be made. For example, extrapolation typically requires that a random variable be assumed to come from a particular shape of distribution e.g. uniform, normal or Weibull. A poor choice of assumed distribution will cause misleading results. Errors from misleading assumptions about distributions have been particularly noted in finance because of the preponderance of hard-to-fit power law distributions there (ref. 22), and there is concern that most complex networked systems have similar problems.

A second problem is lack of data – in some cases, there will not be enough past data to make a model about which we can be confident; there is the risk that a few extreme values will distort the model (either by shifting the numerical values or by giving a misleading impression of the shape of the distribution). This is particularly prevalent in safety, when accidents are rare and hence few data points are available to base statistical models on.

Outside of engineering, statistical methods are a key part of scientific research. It is, of course, the norm for scientific theories to be evaluated by experiments where the results are assessed in statistical terms. However, even in this field (where one might expect higher standards to hold than in engineering and forecasting, given the emphasis on results with long-term, general-purpose validity), it has been suggested that a great many published, peer-reviewed results are false – the experiments do not show what they purport to show (see Ioannidis in (ref. 23) for a recent review of this issue, and some discussion of the reasons behind it).

It can be observed that, since models derived from statistical information do not directly embody the structure of the system or its environment, there is no straightforward way to update a statistical model when the system changes. In this respect, statistical evidence is weaker than any of the other evidence sources discussed in this paper.

Real-world Experimentation and Field Tests

It is the real world trial where reality bites. Bench tests in a laboratory, test flights in empty airspace, and elaborate military field exercises all provide means to test a version of the real system in an approximately real environment. The transition from abstract paper or mental models to a working physical embodiment is a classic time for unexpected flaws to surface. There are a wealth of examples of this, such as Brooks in robotics (ref. 24).

Real-world experimentation is particularly powerful because it can explore both small- and large-scale causes and effects, including all the behaviours that emerge from the full-fidelity interactions of the components. Examples of small-scale causes include the effect of terrain on vehicles, such as a lack of traction on grass or a tendency for sand and soil to get into critical components. Examples of large-scale effects are the behaviour of the actual communications equipment and protocols used in a real distributed data fusion system, given real lines of sight and weather effects.

In safety, real-world experiments have an important role in hazard identification – a hazard exhibited in the live system is much more compelling, and hence harder to ignore, than one derived from an abstract model. Of course, by the time a project reaches real-world test, fixing a hazard is likely to be very expensive. Worse still, hazards revealed in real-world test may lead to real accidents *during the test*.

No real-world experiment or test, however, is the same as actual operational use. This is particularly true in the military sphere, where a field exercise will not feature a genuine enemy force firing with lethal intent, nor real civilians to be caught in the crossfire. Typically, they will also differ in other ways: force sizes, the extent of air support or electronic warfare, and the state of equipment repair. Non-military exercises also differ: empty skies, test tracks, a well-ventilated workshop rather than a cramped engine compartment. These differences are unavoidable, given the cost and risk involved. It means, however, that the test is still a proxy for real usage – it is a *model* of real operation, a *simulation* (although not a *computer* simulation based on an abstracted model).

(There is a further point here – even behaviour in a real operational deployment in an expected operational environment cannot be used to *uncritically* and indiscriminately draw conclusions about behaviour in future deployments. This is, however, outside the scope of this paper.)

A second problem, which is perhaps more serious, is that time and cost constraints severely limit the amount of real-world testing that can be performed. For a complex system (particularly a sociotechnical one where large numbers of humans are part of the system under analysis), it will be practical to explore only a small proportion of its configurations, missions in a subset of its possible operating environments. For each test or exercise, it will only be possible to perform a relatively small number of repetitions compared to that possible with (for example) non-interactive simulation models.

In order to make optimal use of the limited trials that are available, the choice of the conditions of each trial must be heavily prioritised. It is rarely, however, obvious how this should be done. There is a large literature on Design of Experiment techniques, but these inevitably rely on yet another (usually implicit) model of the system, embedding assumptions about its behaviour, in order to guide the process. It is particularly difficult to suggest how to select trials so as to optimally reveal unexpected hazards and problems.

It was noted above that multi-agent simulation is often both epistemically opaque (it is difficult to explain why a certain behaviour occurred) and highly contingent (the behaviour in the late stages of a simulation run may be highly dependent on chance factors early in the run). If such simulations are opaque, however, the real world is worse. A real system, especially a sociotechnical one, is far more complex than any simulation model. The real world is also lamentably lacking in good logging and replay facilities. Indeed, it lacks even the basic facility to (truly) pause a running experiment, or to set a known random seed.

As with the other methods discussed in this paper, real-world tests and exercises are vulnerable to political interference. Allegations of unrealistic conditions in field tests are not uncommon (e.g. Spencer in (ref. 25) on unrealistic restrictions of red force electronic warfare).

Some Conclusions

We can conclude that the several means of safety analysis (and hence sources of safety case evidence) presented here have both strengths and weaknesses. For example, the relative cost of performing experiments means that simulation can explore system behaviour in a *broad* range of situations, whereas real-world testing can provide *deep* insight into precise system behaviour in a smaller number of cases. We can therefore also conclude that each source has value, but *finite* value – there is no source that (in a practical situation rather than a theoretically ideal one) can entirely replace the others. The way forward lies in using the different methods in complementary ways. Given space limitations in this paper, discussions of how we achieve that will have to wait for other venues.

There is a further critical issue. Given that the world is not static, and that we know that any method of analysis will be at best partially successful, we cannot rely exclusively on prior-to-deployment safety analyses. We need to take ongoing account of the new knowledge we gain when our predictions prove to be wrong.

Safety Management as the World Changes

If the world changes, any model based on its previous configuration may be invalidated. When a change is made to an engineered system, when its operating procedures change (officially or in practice), or when important new phenomena appear in its environment, all previous models of that system become suspect. All previous predictions, therefore, become suspect.

There is a second, related challenge: our models are wrong (as Box once put it: *"All models are wrong, some models are useful"* (ref. 26)). As a system is operated, events inevitably occur that reveal omissions, misunderstandings, and unsound extrapolations in the original safety analysis. Although it is of course important that safety engineers make their best effort to provide a predictive safety model ahead of time, the ethical safety engineer must recognise that their modelling and analysis is fallible, and consequently take steps to protect themselves against it.

Both of these problems can be dealt with by updating models (of all kinds) over time based on the actual behaviour of the deployed system. A failure to update a model after reality has invalidated it is a dereliction of duty on the part of the modeller.

Taleb, in (ref. 27) provides a clear example of such a failure in the field of finance. After a stock market hedge fund suffered a massive loss which had been wholly unpredicted by their supposedly state-of-the-art economic models, the fund's econometricians offered the claimed that the loss was a "ten sigma event" (an event of such probability should, according to Taleb, occur once *"every several times the history of the universe"*). Taleb observes that the loss was only "ten sigma" in terms of the previous model – the actual occurrence of the loss has called their model into question, suggesting that such events are much more likely than previously thought (Taleb also suggests that the probability of a truly "ten sigma" event occurring is much less than the probability that the model is wrong).

The traditional source of model invalidation is accidents. Accidents are valuable, but in today's culture, this is no longer acceptable – we demand that accidents be foreseen. (The classic illustration of this is the response of the media to an accident – precursor incidents are found and paraded). Indeed, the recently-issued ICAO[2] Safety

[2] International Civil Aviation Organisation

Management Manual (ref. 28), which provides a general set of rules for safety in international aviation, suggests that the only way to reduce air accident rates further will be to take a 'proactive' approach to safety engineering that actively seeks out areas where safety could be improved (as opposed to a 'reactive' one which responds only to accidents that have occurred).

In order to be proactive, safety-related models (and the analysis based on those models) must be updated based on non-accident data. The most salient source of this is near-miss incidents. (The ICAO manual also requires that general "safety performance measurement" be performed). Such incidents are likely to be far more common than actual accidents, and hence provide a much more useful source of data (one study of industrial accidents produced a '1:600 rule' – for each serious accident that occurred, 600 similar no-loss incidents were reported (ref. 29)). It is easy, however, for an organisation to ignore no-loss (or minor loss) incidents, so efforts must be made to ensure that they are reported, collected, and reviewed.

It follows from the above that the initial safety analysis of a system is only the start. Likewise, the safety case that is built based on that analysis and the certification that is achieved on the grounds of that safety case are only the first steps in ensuring a safe system. They are entry tickets – they allow a system to enter the process of being modelled and analysed in a way that allows its (conditional) safety to be predicted and hence allows it to be operated safely. Safety modelling and analysis is, therefore, a whole-lifetime activity.

Some Implications – "No Epistemic Excuses"

It can be observed that none of the sources discussed earlier provided perfectly accurate predictions – it is the nature of the complex, messy and intractable world we live in that no such predictions are possible. All those approaches, therefore, have the status of *heuristics* – they can be shown to provide useful predictions some of the time. It is incumbent on the user of such heuristics to know their capabilities and limitations, and to caveat their predictions accordingly.

It follows from the above that in safety analysis we are never in a position of objective knowledge; we never possess objective evidence of safety. The UK military safety standard Def Stan 00-56 Issue 4 may state that (when building a safety case) "*Where possible, objective evidence shall be provided.*" (ref. 6), but a safety assessor expecting such evidence will be disappointed.

A second corollary is that there is no escape from normative epistemic standards. The use of explicit safety cases (as opposed to solely process-based certification) makes this extremely clear. A safety engineer must predict, and they must decide *how* they are going to predict. They will need to justify their choice of prediction methods, given both the practical history of the techniques available and projections of the validity of new techniques (or of techniques that have been used in non-safety domains). Given the complexity of the world we inhabit, empirical evaluation of prediction techniques is important, although the nature of safety engineering is that we deal with very-low-probability events, which makes true empirical evaluation extremely difficult. It is through shared norms, shared standards of evidence, that product-based certification is possible.

This reliance on explicit prediction raises a number of challenges. One example – it may be difficult to establish epistemic norms that allow us predict what we already can achieve. For example, McDermid and Kelly note in (ref. 30) that current aviation software seems to achieve a dangerous failure rate of around 10^{-7} per year, but that for a given software product we cannot provide ahead-of-time evidence that the dangerous failure rate will be better than 10^{-4} per year.

One final observation can be made. Once we introduce the concept of epistemic norms, whereby we are providing certain rules for the techniques and the evidence that can be used to make certain claims, we are providing means by which we delegate parts of the understanding and decision-making process to explicit procedures. Such explicit procedures can be implemented by machines (this is particularly true for simulation). This is directly contiguous with the endeavour of Artificial Intelligence. In our case, however, we are describing ways by which humans and machines can work *together*, a less ambitious aim but one that is far more promising in the short to medium term.

Conclusions

It is the contention of this paper that simulation evidence for safety is as good (or as bad) as it proves to be in practice – that the question as to the validity of simulation results in a safety case is primarily an empirical matter. There is a need for generalised conclusions drawn from empirical data as to how effective simulations are at predicting system safety (and, within that, how effective different simulation formalisms and approaches are in different situations). These conclusions will provide the epistemic norms which dictate what safety case claims can be supported by what simulation evidence. It follows that there is a need for them to be condensed into guidance for safety case assessors.

This is true, however, for *all* forms of prediction and all types of evidence. Engineering intuition, statistics from past activity, and real-world trials are all imperfect sources of knowledge. It follows that we must hold all sources of evidence to the same standard. In particular, we must understand how our available techniques can be used together in a complementary way, where the strengths of one cover for the weaknesses of another.

There is an ethical obligation for safety engineers to make predictions and to make decisions accordingly. The public look to safety professionals to prevent accidents, to prevent accidental material damage and loss of life. An engineer who cannot predict does not know, and hence cannot ethically make claims about system safety. There is also an ethical obligation to make predictions using the best methods available, and this translates into an obligation for the safety engineering community to identify those methods and the role that each has. To reject techniques (such as simulation) based on anecdotes or prejudice is unacceptable.

References

[1] H. A. P. Blom, S. H. Stroeve, and H. H. d. Jong, "Safety Risk Assessment by Monte Carlo Simulation of Complex Safety Critical Operations," in Proceedings of the Fourteenth Safety-critical Systems Symposium, Bristol, UK, 2006.

[2] C. Johnson, "The glasgow-hospital evacuation simulator: Using computer simulations to support a risk-based approach to hospital evacuation," Technical report, University of Glasgow 2005.

[3] R. Alexander, D. Kazakov, and T. Kelly, "System of Systems Hazard Analysis Using Simulation and Machine Learning," in Proceedings of the 25th International Conference on Computer Safety, Reliability and Security (SAFECOMP'06), 2006.

[4] K. A. Richardson, "'Methodological implications of complex systems approaches to sociality': some further remarks," *Journal of Artificial Societies and Social Simulation*, vol. 5, pp. 1-9, 2002.

[5] E. S. Ferguson, "How Engineers Lose Touch," *Invention and Technology*, pp. 16-24, 1993.

[6] "MoD Interim Defence Standard 00-56 Issue 4 - Safety Management Requirements for Defence Systems," Ministry of Defence 2007.

[7] V. L. Arntsen, "Summation of risk: Assessment of total system risk for complex systems," Uppsala University 2007.

[8] "MIL-STD-882E - standard practice for system safety (draft)," Department of Defense 2006.

[9] A. M. Law, *Simulation Modeling and Analysis*: McGraw-Hill Higher Education, 2006.

[10] A. Ilachinski, "Exploring self-organized emergence in an agent-based synthetic warfare lab," *Kybernetes: The International Journal of Systems & Cybernetics*, vol. 32, pp. 38–76, 2003.

[11] P. Humphreys, *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press, 2004.

[12] P. Roman, "Garbage In, Hollywood Out!," in *SimTecT 2005*, 2005.

[13] K. A. Richardson, G. Mathieson, and P. Cilliers, "The Theory and Practice of Complexity Science-Epistemological Considerations for Military Operational Analysis," in Proceedings of SysteMexico, 2000.

[14] "Columbia Accident Investigation Board, “Report volume 1” " National Aeronautics and Space Administration 2003.

[15] N. Oreskes and K. Belitz, "Philosophical issues in model assessment," *Model Validation: Perspectives in Hydrological Science*, pp. 23-41, 2001.

[16] J. Suokas and R. Kakko, "On the problems and future of safety and risk analysis," *J. HAZARDOUS MATER.*, vol. 21, pp. 105-124, 1989.

[17] R. M. Dawes and R. Hastie, "Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making," Sage Publications, 2001.

[18] P. E. Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*: University of Minnesota Press, 1954.

[19] W. M. Grove and P. E. Meehl, "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy," *Psychology, Public Policy, and Law*, vol. 2, pp. 293-323, 1996.

[20] J. Shanteau, "Competence in experts: The role of task characteristics," *Organizational Behavior and Human Decision Processes*, vol. 53, pp. 252-266, 1992.

[21] M. Resnick, "Beyond the Centralized Mindset," *Journal of the Learning Sciences*, vol. 5, pp. 1-22, 1996.

[22] N. Taleb, *The Black Swan: The Impact of the Highly Improbable*: Allen Lane, 2007.

[23] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, 2005.

[24] R. A. Brooks, "Artificial life and real robots," *Proceedings of the First European Conference on Artificial Life*, pp. 3-10, 1992.

[25] H. Spencer, "Re: "Computer Models Leave U.S. Leaders Sure of Victory"," 1991.

[26] G. E. P. Box, "Robustness in the strategy of scientific model building," *Robustness in Statistics*, pp. 201-236, 1979.

[27] N. Taleb, "Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets," New York: Random House, 2005.

[28] "Safety Management Manual," International Civil Aviation Organization 9859, 2006.

[29] F. E. Bird and G. L. Germain, *Practical Loss Control Leadership*: International Loss Control Institute, 1985.

[30] J. McDermid and T. Kelly, "Software in Safety Critical Systems: Achievement and Prediction," *Nuclear Future*, vol. 2, pp. 140-146, 2006.

Biography

Dr Robert Alexander, Ph.D., Department of Computer Science, University of York, Heslington, York, YO10 5DD, UK, telephone – +44 1904 432792, facsimile – +44 1904 432767, e-mail – robert.alexander@cs.york.ac.uk

Dr Robert Alexander is a Research Associate in the High Integrity Systems Engineering (HISE) group in the Department of Computer Science at the University of York. Since October 2002 he has been working on methods of safety analysis for systems of systems and autonomous systems, with a particular emphasis on simulation and automated analysis. Robert graduated from Keele University in 2001 with first class honours in Computer Science, and was awarded his doctorate in 2008 by the University of York.

Dr Tim Kelly, Ph.D., Department of Computer Science, University of York, Heslington, York, YO10 5DD, UK, telephone – +44 1904 432764, facsimile – +44 1904 432708, e-mail – tim.kelly@cs.york.ac.uk

Dr Tim Kelly is a Senior Lecturer in software and safety engineering within the Department of Computer Science at the University of York. He is also Deputy Director of the Rolls-Royce Systems and Software Engineering University Technology Centre (UTC) at York. His expertise lies predominantly in the areas of safety case development and management. His doctoral research focused upon safety argument presentation, maintenance, and reuse using the Goal Structuring Notation (GSN). Tim has provided extensive consultative and facilitative support in the production of acceptable safety cases for companies from the medical, aerospace, railways and power generation sectors. Before commencing his work in the field of safety engineering, Tim graduated with first class honours in Computer Science from the University of Cambridge. He has published a number of papers on safety case development in international journals and conferences and has been an invited panel speaker on software safety issues.