Senate Inquiry Supplementary Submission on the ATSB Findings of the Ditching of the Westwind II VH-NGA off Norfolk Island on 18 November 2009

This fifth submission is made under Parliamentary Privilege 18 March 2013

by Bryan Aherne in a private capacity

Contents

1. Introduction

This submission is aimed at contemporary research into human factors, human error and safety systems, and the ATSB use of the IPCC (Intergovernmental panel on climate change) verbal probabilities in their internal risk methodology process.

The purpose of this submission is so the committee may consider what other experts say about risk communication using the IPCC methodology, human factors, human error and safety systems in the context of complex socio-technical systems.

2. ATSB use of IPCC specific language to describe risk and probability in their methodology.

In 2007, the ATSB adopted the definitions of probability in its analysis activities and for investigation reports from the Intergovernmental Panel on Climate Change (IPCC) . The IPCC in 2005 developed a standardised terminology to facilitate communication of uncertainty of technical information in its field. The IPCC definitions were a result of extensive and considerable rigour by international experts (ATSB, 2008).

The research of Patt & Schrag (2003) found that the IPCC Third Assessment Report (AR 3) used words differently from the way lay readers of assessments do. Experiments on undergraduate science students confirmed that the language of the IPCC authors to describe uncertainty, depended only on the probability of the outcome or certainty which they believed would occur, and not on the events magnitude. The language that lay people use to discuss uncertainty and the meanings of the associated descriptors are event dependant and context driven, that is, the events total risk (Patt & Schrag, 2003). pp 28.

When both the communicator and audience are using uncertainty descriptors to describe risk, and *not probability*, accurate understanding passes from communicator to audience without bias (see Brun & Teigen, 1988 in Patt & Schrag, 2003). pp 23.

However, when communicators use verbal probabilities, the audience still interprets them as describing risk and thus, result in miscommunication. The results mean an audience can underweight probability of high magnitude events. It also suggests that probability descriptors are used and interpreted by people as containing information about event magnitude. People are also more likely to choose more certain sounding probability descriptors (eg., *likely* instead of *unlikely*) to discuss increasing severity consequences, and expect a certain amount of exaggeration about likelihood of severe consequence events (Patt & Schrag, 2003). pp 26.

Patt & Schrag believed the use by the IPCC of a fixed scale verbal probability descriptor could introduce unintended bias of under-responding to aggregate risks. This means that the reader (or viewer) would need to scrutinise and carefully read the report. However, in the case where a report is not read carefully, bias enters when readers make intuitive judgements about likelihood when they fail to match verbal probabilities with probability ranges (Patt & Schrag, 2003). pp 29.

The observation of Patt & Schrag (2003) regarding the risks of introducing intuitive judgement when using fixed scale of verbal probabilities seems to support the bias in scientific fields, such as earth science.

The IPCC changed two elements of the verbal description of '*Extremely unlikely'* and '*Medium Likelihood'* to '*Exceptionally unlikely'* and' *About as likely as not'* respectively in AR 3 to AR 4.

In AR 4,(2006) the IPCC codified linguistic terms to avoid ambiguity and introduced a "level of confidence" table based on the degree of understanding within an expert community when data is lacking or incomplete and where subjective judgement is required (Risbey & Kandlikar, 2007). pp 20. However, Risbey & Kandlikar concluded that the "level of confidence" table should only be used to condition the form in which likelihood is expressed, rather than the value itself (Risbey & Kandlikar, 2007). pp 30.

I cannot find in the ATSB methodology their use of the "level of confidence" table as the IPCC introduced in 2006 (Risbey & Kandlikar, 2007) to validate their confidence of evidence. They appear to use their existence, importance and influence technique.

If the ATSB methodology is unable to ascertain the most basic of safety issues, (see my third submission of the most obvious failed or absent risk controls), then this may require a review of the workability and validity of this ATSB existence, influence and importance approach to identification of safety issues.

Identification of such issues is the prime responsibility of the ATSB under the TSI Act. ATSB have failed to recognise at least nine serious safety issues identified in the submissions to the committee.

3. Some contemporary research on human factors, human error, human reliability and systems safety.

French, et al, (2011) primarily wrote this paper in addressing the limitations of human reliability analysis (HRA) as a method used in quantitative risk assessment (QRA), however, the authors never the less provide useful insight into current limitations and contemporary thinking into the areas of human factors, human error, reliability and systems safety.

"A key issue is that HRA focuses on human errors, whereas many systems failures may arise not just despite, but sometimes because of fully appropriate and rational behaviour on the part of those involved. Thus we need a broader understanding of human behaviour than that relating to human error. We also need to recognise that cultural, organisational, social and other contexts influence behaviour, perhaps correlating behaviour across a system, thus invalidating assumptions of independence commonly made in risk and reliability analyses. One of the flaws common to many current HRA methodologies is that they tend to focus on easily describable, sequential, generally low-level operational tasks. Yet the human behaviour that is implicated in many system failures may occur in other quite different contexts, maybe in developing higher level strategy or during the response to an unanticipated initiating failure event." French, et al, 2011, page 755

"Much of the discussion around these models focussed on the issue that errors of omission (failures to respond to events appropriately) were considered easier to model than errors of commission, i.e. inappropriate human actions. However, this simplistic dichotomy now appears too stark in light of our current, richer, qualitative understandings of human cognition, motivation and decision making, including the effects of stress, emotion, training, group interactions, organisational structures, cultures and so forth (Bazerman, 1999, 2006; French et al., 2009; Kahneman et al., 1982; Kahneman and Tversky, 2000). Research in these fields has

shown that there are systematic influences on decision making and behaviour that cannot be categorised simply as omissions or commissions": French, et al, 2011, page 756

"But human behaviour can correlate the risks of failure of two or more barriers, and most systems also harbour the opportunity for the 'bypass' of these barriers. Human behaviour and propensity to failure varies in complex ways with, e.g., their tiredness, stress and general emotional state, which may well be influenced by external events leading to a common cause and which may disrupt several safety barriers simultaneously. For instance, the Chernobyl Accident (International Atomic Energy Agency, 1991; Marples 1997) was in large measure caused by the imperative to conduct an engineering experiment within a fixed time, leading to stress in the operators and behaviour that compromised several of the safety barriers simultaneously. Another potential unsafe behaviour is to discover an indication of a 'hole' in one layer and to defer further investigation, relying on the 'cover' offered by other layers: such behaviour occurred during a recent leak of radioactivity at Sellafield (Adhikari et al., 2008). Hrudey et al. (2006) describe similar behaviour during the Walkerton drinking water tragedy in Ontario, where latent and active flaws left unaddressed exacerbated the impact of agricultural run-off infiltrating a town's shallow groundwater supply. On the positive side, humans have the ability to recover, to respond to the unexpected, to think 'out of the box', and so on, effectively repairing a compromised layer or even introducing a new one – the latter is, of course, the principle of preventative risk management." French, et al, 2011, page 756

"Errors are just one of a range of behavioural products of a number of individual and organisational precursors; they are not a class of behaviours that are entirely distinct from other behaviours and thus should not be considered in isolation. In the organisational context, it is often an external system or judgement that categorises a behaviour as an error rather than the behaviour itself being inherently and indisputably wrong." French, et al, 2011, page 757

"Real human judgement and decision making is not as rational and analytic as one might wish. Since the early 1980s, psychologists have distinguished between two different forms of thinking (Chaiken et al., 1989)3:

_ System 1 thinking, often referred to as 'intuition' or 'gut reaction' that involves a superficial analysis/interpretation of the relevant information based on much simpler forms of thinking on the fringes or outside of consciousness.

_ System 2 thinking, characterised by conscious analytical thought that involves a detailed evaluation of a broad range of information, often based on a rule that is assumed to provide the 'correct' answer or solution."

"While formal risk assessment techniques have the characteristics of System 2 thinking, system operators may use System 1 thinking in their day-to-day operations and responses to events. For example, a nuclear power plant is the outcome of considerable complex analysis, research and design, i.e. System 2 thinking. The operators of such a plant, however, do not typically engage in the same kind of analytical thinking as the system engineers and designers. The operators' work comprises much more routine procedures and, where complex problems are faced, there is potential for operators to make them more manageable through System 1 heuristics. It has become common to refer to much of System 1 thinking as involving 'heuristics and biases', because of its deviation from the more rational, analytic System 2 thinking, though that terminology is as pejorative as the constant use of the term 'human error' in HRA which we reject in this article."

"It is of concern that very little use of this extensive, often empirically based literature has been made in developing HRA methodologies. Indeed, the mechanistic approach common to many such methodologies based on fault tree representations of human action assumes that the operators are using System 2 thinking when in all probability their intuitive responses and actions are guided by System 1 thinking (Bargh et al., 1996)."

This recognition that human behaviour is complex and driven by a range of internal and external factors leads us to question the value of terminology such as 'error', 'slip' or 'failure' within HRA. Human errors and faults are socially defined events: a perfectly reasonable action to one person may be an unreasonable failure to another (Hollnagel, 2000). Furthermore, however well judged a decision may be a priori, it may through 'ill fortune' lead to unwanted outcomes. Hence what may seem an error in hindsight may not be the outcome of irrational or erroneous choice. We should focus more on human behaviour in individual, group and organisational contexts and recognise its potential involvement in system failure – without the pejorative judgement of whether that behaviour is aberrant in any sense. For example, in the Three Mile Island Incident (Commission on the Three Mile Island Accident, 1979) the initiating event – the formation of a hydrogen bubble which forced down cooling water exposing the core – had not been anticipated in the reactor's design or safety studies. The operators not only did not recognise what was happening, but also had never anticipated that it might. It was an incident beyond their experience and imagination, in a very real sense outside of scientific and engineering knowledge as it stood then. The operators behaved entirely sensibly and in accordance with their mental models of what they believed was happening. There was no error in their behaviour in this respect, not at least in the sense of human error within HRA theory As we build and operate more and more complex systems, we should recognise that it is inevitable that we will encounter unanticipated events and conditions. Risk and reliability analyses need to take account of human responses to these and, although those responses

may indeed lead to untoward outcomes, it is far from clear
that they should be dubbed errors. " French, et al, 2011, page 757, 758.

Of particular note in the above work, is that of the Systems 1 & 2 thinking. It may be that the ATSB SIIMS methodology, is using a Systems 2 type approach to its analysis, without consideration of the Systems 1 thinking of the operational persons involved in transport accidents and incidents. I have not used the SIIMS methodology so am unable to comment on that.

However it seems that proven basic safety investigation methods have been disentangled in favour of a pure academic approach.

If this is the case, the pure academic approach needs validation and scrutiny by peer experts in human factors, safety investigation and organisational systems safety.

I hope this information and the attached research papers can assist the committee in its work.

Yours sincerely,

Bryan Aherne

Safety Science 49 (2011) 753–763

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Review

Human reliability analysis: A critique and review for managers

Simon French [a,*], Tim Bedford [c], Simon J.T. Pollard [b], Emma Soane [d]

[a] Manchester Business School, University of Manchester, Manchester M15 6PB, UK
[b] The Collaborative Centre of Excellence in Understanding and Managing Natural and Environmental Risks, Cranfield University, Cranfield MK43 0AL, UK
[c] Department of Management Science, University of Strathclyde, Glasgow G1 1QE, UK
[d] Employment Relations and Organisational Behaviour Group, Department of Management, London School of Economics and Political Science, London WC2A 2AE, UK

ARTICLE INFO

Article history:
Received 11 December 2009
Received in revised form 7 December 2010
Accepted 14 February 2011
Available online 21 March 2011

Keywords:
Cynefin model of decision contexts
High reliability organisations
Human reliability analysis (HRA)
Management of risk

ABSTRACT

In running our increasingly complex business systems, formal risk analyses and risk management techniques are becoming more important part to managers: all managers, not just those charged with risk management. It is also becoming apparent that human behaviour is often a root or significant contributing cause of system failure. This latter observation is not novel; for more than 30 years it has been recognised that the role of human operations in safety critical systems is so important that they should be explicitly modelled as part of the risk assessment of plant operations. This has led to the development of a range of methods under the general heading of *human reliability analysis* (HRA) to account for the effects of human error in risk and reliability analysis. The modelling approaches used in HRA, however, tend to be focussed on easily describable sequential, generally low-level tasks, which are not the main source of systemic errors. Moreover, they focus on errors rather than the effects of all forms of human behaviour. In this paper we review and discuss HRA methodologies, arguing that there is a need for considerable further research and development before they meet the needs of modern risk and reliability analyses and are able to provide managers with the guidance they need to manage complex systems safely. We provide some suggestions for how work in this area should develop. But above all we seek to make the management community fully aware of assumptions implicit in human reliability analysis and its limitations.

Contents

1. Introduction

Complex systems are never 100% reliable: they fail, sometimes catastrophically, more usually reparably. Perrow (1984, 1994) has argued that failures are an inevitable consequence of the increasing complexity of our systems. Whatever the case, inevitable or not, failures undoubtedly occur. Even in systems that appear to be largely technological rather than human, we find that in the majority of cases there is a human element involved. Maybe some erroneous or even malicious behaviour initiates the failure; maybe the human response to some event is insufficient to avoid system failure; or maybe the original design of the system did not anticipate a potential failure or unfavourable operating conditions.

Statistics show human error is implicated in (see also Hollnagel, 1993):

* Corresponding author. Address: Manchester Business School, Booth Street West, Manchester M15 6PB, UK.
E-mail address: simon.french@mbs.ac.uk (S. French).

0925-7535/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ssci.2011.02.008

- over 90% of failures in the nuclear industry (Reason, 1990a), see also (United States Nuclear Regulatory Commission, 2002);
- over 80% of failures in the chemical and petro-chemical industries (Kariuki and Lowe, 2007);
- over 75% of marine casualties (Ren et al., 2008);
- over 70% of aviation accidents (Helmreich, 2000);
- over 75% of failures in drinking water distribution and hygiene (Wu et al., 2009).

In addition to highly technological industries, there are other complex systems involving applications of technology in which we include complex mathematical modelling, software and web-based systems. The growth of service industries with new business models implies an even greater dependence of businesses, organisations and even economies on reliable human interactions. For instance, recently human checks and balances failed to detect dubious investment behaviour of a trader at *Société Générale* and led to a loss of some €4.9bn, large enough to have economic and financial effects beyond the bank. The current 'credit crunch' owes not a little to misjudgement and error in the banking and finance sectors, indicating the growing interdependence of many disparate parts of the modern global economy. It also owes a lot to a loss of investors' confidence and trust, both of which inform human behaviour. These data indicate how vulnerable our systems are, even after many years of refinement and improvement; and how important an understanding of human behaviour is if we are to reduce the risk to systems. Another high profile example is the leak in the THORP plant at Sellafield (Thermal Oxide Reprocessing Plant) that was discovered in 2005 (see Board of Inquiry, 2005). This relatively modern plant had been designed to a high standard of safety, but information indicating a system problem was available for some months and yet went unnoticed. Despite previous incidents in 1998 and earlier in 2005, the information that should have suggested a leak, or at least a problem requiring investigation, was misinterpreted. The prevailing attitude was that the system was error-free and hence information that could suggest the contrary was ignored or dismissed.

Managerial processes are critical to successful operation of any complex system; and the quality of management processes depends on their understanding of the import and limitations of the results of the risk (and other) analyses that are provided to them. We emphasise here that all managers, whether or not they have an explicit responsibility for risk management, need to have some understanding of the assumptions and limitations of such analyses. In this article, we examine current and past approaches to human reliability analysis (HRA). We discuss its assumptions, limitations and potential in qualitative terms so that managers can better assess the value of the information that it provides them and so manage risks more effectively. We also suggest that further development of HRA methodologies should take more account of the managerial practices that could be applied to reduce the failures that occur at the interface of human behaviour and technology.

Managers understand human behaviour; good managers understand human behaviour extremely well. To bring out the best in a team one needs to know how each will respond to a request, an instruction, an incentive or a sanction. Yet only the most foolhardy and overconfident of managers would claim that they can predict human behaviour perfectly all the time – or even 95% of the time. The problem is that we often need to design systems with very high reliabilities, many times with overall failure rates of less than 1 in 10 million (i.e. 1 in 10^{-7}). To design and analyse such systems we need a deep understanding of human behaviour in *all* possible circumstances that may arise in their management and operation. And that is the challenge facing HRA. Our current understanding of human behaviour is not sufficiently comprehensive: worse, current HRA methodologies seldom use all the understanding that we do have.

Of course, there is a trivial mathematical answer to this. If we are to achieve an overall system reliability of 10^{-7}, we do not need humans to be perfectly reliable. We simply need to know how reliable they are and then ensure that we arrange and maintain sufficient safety barriers around the system to ensure that overall system failure probabilities are as low as required. Suppose we construct seven independent safety barriers perhaps some involving humans, some purely technological and suppose each has a probability of 1 in 10 of failing, then arranging them (conceptually) in sequence so that the whole system fails if and only if every one of the seven fails gives an overall probability of system failure of

$$\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}\times\frac{1}{10}=10^{-7}.$$

The problem with this is that there are few barriers that are truly independent, most systems offer opportunities to 'bypass' these barriers. Moreover, human behaviour tends to introduce significant correlations and dependencies which invalidate such calculations, reducing the benefit that each extra safety barrier brings; such problems with protective redundancy are well known (for example, Sagan, 2004). So the simplistic calculation does not apply, and we shall argue that we have yet to develop sufficiently complex mathematical modelling techniques to describe human behaviour adequately for risk and reliability analyses.

In many ways the roles of risk and reliability analysis in general and of HRA in particular are often misunderstood by system designers, managers and regulators. In a sense they believe in the models and the resulting numbers too much and fail to recognise the potential for unmodelled and possibly unanticipated behaviours – physical or human – to lead to overall system breakdown (cf. French and Niculae, 2005). Broadly there are two ways in which such analyses may be used.

- When HRA is incorporated into a *summative* analysis, its role is to help estimate the overall failure probabilities in order to support decisions on, e.g., adoption, licensing or maintenance. Such uses require quantitative modelling of human reliability; and overconfidence in these models can lead to overconfidence in the estimated probabilities and poor appreciation of the overall system risks.
- There are also *formative* uses of HRA in which recognising and roughly ranking the potential for human error can help improve the design of the system itself and also the organisational structures and processes by which it is operated. Effective HRA not only complements sound technical risk analysis of the physical systems, but also helps organisations develop their safety culture and manage their overall risk. Indeed, arguably it is through this that HRA achieves its greatest effect.

These uses are not independent – in designing, licensing and managing a system one inevitably iterates between the two – they do differ, however, fundamentally in philosophy. In summative analysis the world outside the system in question learns from the outcome of an analysis; in formative analysis the world inside the system learns from the process of analysis. In summative analysis the ideal is almost to be able to throw away the process and deal only with the outcome; in formative analysis the ideal is almost to throw away the outcome and draw only from the process. While we believe that HRA has a significant potential to be used more in formative ways; we are concerned at its current ability to fulfil a summative role, providing valid probabilities of sequences of failure events in which human behaviour plays a significant role. We believe that there is scope for considerable overconfidence in the summative

power of HRA currently and that management, regulators and society in general need to appreciate this, lest they make poorly founded decisions on regulating, licensing and managing systems.

The four of us were part of a recent UK EPSRC funded multi-disciplinary project *Rethinking Human Reliability Analysis Methodologies* to survey and critique HRA methodologies (Adhikari et al., 2008). Our purpose in this paper is to draw out the relevant conclusions from this project for the management community and, perhaps as well, for our political masters who create the regulatory context in which complex systems have to operate. Overall we believe that current practices in and uses of HRA are insufficient for the complexities of modern society. We argue that the summative outputs of risk and reliability analyses should be taken with the proverbial pinch of salt. But not all our conclusions will be negative. There is much to be gained from the formative use of HRA to shape management practices and culture within organisations and society which can lead to better, safer and less risky operations.

In the next section we briefly survey the historical development underlying concepts of HRA and its role in risk and reliability analyses. We reflect on the widely quoted Swiss Cheese Model (Reason, 1990b), which seeks to offer a qualitative understanding of system failure – though we shall argue that it may actually lead to systematic misunderstandings! In Section 0 we turn to modern theories of human behaviour, particularly those related to judgement and decision. A key issue is that HRA focuses on human *errors*, whereas many systems failures may arise not just *despite*, but sometimes *because of* fully appropriate and rational behaviour on the part of those involved. Thus we need a broader understanding of human behaviour than that relating to human error. We also need to recognise that cultural, organisational, social and other contexts influence behaviour, perhaps correlating behaviour across a system, thus invalidating assumptions of independence commonly made in risk and reliability analyses. One of the flaws common to many current HRA methodologies is that they tend to focus on easily describable, sequential, generally low-level operational tasks. Yet the human behaviour that is implicated in many system failures may occur in other quite different contexts, maybe in developing higher level strategy or during the response to an unanticipated initiating failure event. In recent years there have been many studies of organisational forms which seem to be more resilient to system failures than might be expected and we discuss such studies of *high reliability organisations* (HROs) briefly in Section 0. Another flaw common to many current HRA methodologies is the lack of specification of the domain of applicability – hence making it difficult to select appropriate methods for a given problem. Therefore in Section 0, we use Snowden's Cynefin classification of decision contexts (Snowden, 2002; Snowden and Boone, 2007) to categorise different circumstances in which human behaviour may be involved in system failure. We believe that the use of Cynefin – or a similar categorisation of decision contexts – can help in delineating when different HRA methodologies are appropriate. Moreover, it points to areas in which we lack a really sound, appropriate HRA methodology. Our final two sections draw our discussion to a close, suggesting that:

- by drawing together current understandings from HRA with other domains of knowledge in behavioural, management and organisational theories, we can make better formative use of HRA in designing systems, process and the organisations that run these;

but that:

- the state of the art in quantitative HRA is too poor to make the summative assessments of risk and reliability that our regulators assume, and that society urgently needs to recognise this.

2. HRA methodologies and the Swiss cheese model

Reliability analysis and risk analysis are two subjects with a great deal of overlap (Aven, 2003; Barlow and Proschan, 1975; Bedford and Cooke, 2001; Høyland and Rausand, 1994; Melnick and Everitt, 2008). The former is generally narrower in scope and tends to deal with engineered systems subject to repeated failures and the need for preventative maintenance policies to address these. Key concepts in reliability engineering include component availability, reliability and maintainability; mean times to, and between failure; the use of specific fault tree and failure mode tools; and the concepts of system redundancy. Reliability engineering owes a significant amount to advances in manufacturing engineering and the desire to improve production quality and optimise output (Lewis, 1994). Risk analysis is a much broader term and tends to deal with more one-off failures that may write-off a system with concomitant impacts elsewhere. It is not necessarily restricted to technical systems and has developed into a broad interdisciplinary field with important inputs from the social sciences, alongside applied mathematics and decision science. But both reliability engineering and risk analysis are essentially are concerned with anticipating possible failures and assessing their likelihood. HRA specifically relates to methodologies for anticipating and assessing the effect of those failures which relate to human action or inaction, and not the failure of some physical component.

There are many reasons why one might undertake a risk or reliability analysis. In broad terms the first three items in our list relate to formative uses of risk and reliability analysis and the last two to summative uses.[1]

1. The designers of a system may be concerned with 'designing out' the potential for system failure. Part of this involves analysing how human behaviour may affect the system in its potential both to compromise its reliability and to avoid the threat of imminent failure.
2. Sometimes an organisation wants to restructure and change its reporting structures. In such circumstances, it may wish to understand how its organisational design may affect the reliability and safety of its systems; and in turn that understanding may inform the development of its operating practices and safety culture.
3. There may be a need to modify a system in which case there are needs to design the modification *and* the project to deliver the modification.
4. During licensing discussions between a government regulator and the system operator there may be a need to demonstrate that a system meets a safety target. An assessment of the risks arising from human behaviour will be an integral part of this.
5. There may be a need to choose which of several potential systems to purchase and the risk of system failure may be a potential differentiator between the options. Such differences may not be purely technical, since some systems may be more or less at risk from some human behaviours.

As a component of a full risk or reliability analysis, HRA may be used in any of these ways.

The origins of HRA lie in the early probabilistic risk assessments performed as part of the US nuclear energy development programme in the 1960s (Bedford and Cooke, 2001; United States Nuclear Regulatory Commission, 1975) Early first generation HRA methods such as the *Technique for Human Error Rate Prediction* (THERP) (Swain and Guttmann, 1983) were very similar to those in other areas of reliability analysis: namely, the probability of a

[1] We make no claims that this list is exhaustive, just sufficient for our discussion.

human error is assessed via a simple event tree analysis. The event tree simply listed an initiating event', which might be a system error reaching the human operator, and then considered a series of tasks that which had to be correctly carried out to prevent unwanted consequences. Essentially, in these early models, the human operator is treated as another component in the system. Hollnagel (1993) referred to this general approach as decomposition. A variety of other first generation methods have been developed with broadly similar features to THERP – the use of task analysis, use of nominal probabilities for task failure, adjustment factors to take account of different performance conditions, error factors and so on. The *Human Reliability Analysis Event Tree* method (HEART) (Williams, 1985) is a good example of a method that aims to use many of the same features but in a simplified setting to give a more straightforward approach. Recognising that many tasks have an associated time for completion, the *Human Cognitive Reliability* method (HCR) (Hannaman et al., 1984) modelled the time to successful completion. A wider review of these and many other methods is given in Kirwan (1994).

Much of the discussion around these models focussed on the issue that errors of *omission* (failures to respond to events appropriately) were considered easier to model than errors of *commission*, i.e. inappropriate human actions. However, this simplistic dichotomy now appears too stark in light of our current, richer, qualitative understandings of human cognition, motivation and decision making, including the effects of stress, emotion, training, group interactions, organisational structures, cultures and so forth (Bazerman, 1999, 2006; French et al., 2009; Kahneman et al., 1982; Kahneman and Tversky, 2000). Research in these fields has shown that there are systematic influences on decision making and behaviour that cannot be categorised simply as omissions or commissions: see Section 0 below. Human failure is far more complex than the failure of, say, a steel support beam or a hard disk. To be fair, second[2] generation HRA methods (Barriere et al., 2000; Hollnagel, 1993) attempted to incorporate contextual effects such as tiredness, stress and organisational culture on an operator's proneness to error; and third generation HRA methods (Boring, 2007; Mosleh and Chang, 2004) have sought to allow for the potential variation in response and recovery actions once an error chain has begun. Notwithstanding this, we argue that far more development is needed before any method takes account of all our current understandings of human behaviour.

Surveys of current HRA methodologies may be found in Adhikari et al. (2008), Forester et al. (2006) and Hollnagel (1993, 1998). For other recent research and developments in HRA, see the special issue of the *Journal of Loss Prevention in the Process Industries* (2008, 21, 225-343). Software reliability analysis also has a large literature (Courtois et al., 2000; Lyu, 2005; Zhang and Pham, 2000). Software engineering is largely an endeavour of human design and thus subject to all the risks that HRA seeks to explore and assess. To date, software reliability assessment has, by and large, also adopted a mechanistic or empirical modelling of human error similar in methodology to current quantitative HRA.

Reason (1990b) offered a metaphor for system failure involving human error likening failure processes to movements of slices of Swiss cheese relative to each other: see Fig. 1. Essentially this suggested that systems do not fail because of a single failure, but because several elements fail near simultaneously, as if the holes in slices of Swiss cheese have aligned. Although it is clear from his writings that Reason knew the limitations of metaphors (Reason, 1995, 1997), his readers have often interpreted the model too mechanistically. There has been a dominant tendency to imagine

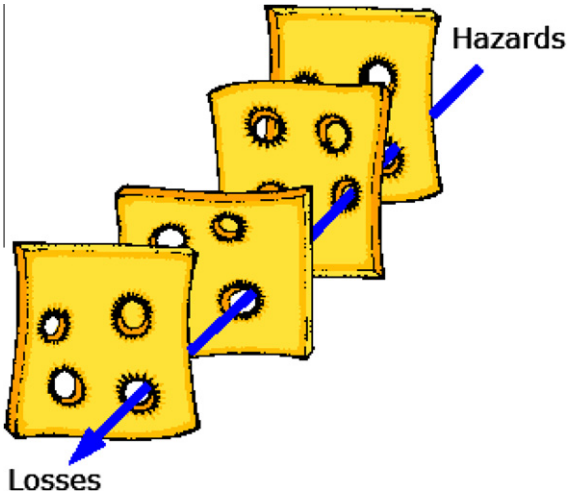

Fig. 1. Reason's Swiss cheese model (Reason, 1990b).

a *fixed* number of slices, sliding backwards and forwards *independently* of each other until a series of holes align. In safety studies one talks of the number of safety barriers (the multi-barrier concept) or layers between normal operation and system failure; and, in a sense, the slices of Swiss Cheese mirror these. Systems are designed with a set number of safety barriers and these barriers are intended to be independent: cf. the simplistic calculation of a failure rate of 1 in 10^7 above. But human behaviour can correlate the risks of failure of two or more barriers, and most systems also harbour the opportunity for the 'bypass' of these barriers. Human behaviour and propensity to failure varies in complex ways with, e.g., their tiredness, stress and general emotional state, which may well be influenced by external events leading to a common cause and which may disrupt several safety barriers simultaneously. For instance, the Chernobyl Accident (International Atomic Energy Agency, 1991; Marples 1997) was in large measure caused by the imperative to conduct an engineering experiment within a fixed time, leading to stress in the operators and behaviour that compromised several of the safety barriers simultaneously. Another potential unsafe behaviour is to discover an indication of a 'hole' in one layer and to defer further investigation, relying on the 'cover' offered by other layers: such behaviour occurred during a recent leak of radioactivity at Sellafield (Adhikari et al., 2008). Hrudey et al. (2006) describe similar behaviour during the Walkerton drinking water tragedy in Ontario, where latent and active flaws left unaddressed exacerbated the impact of agricultural run-off infiltrating a town's shallow groundwater supply. On the positive side, humans have the ability to recover, to respond to the unexpected, to think 'out of the box', and so on, effectively repairing a compromised layer or even introducing a new one – the latter is, of course, the principle of preventative risk management.

In terms of the Swiss Cheese model, many of these failings correspond to varying the size of the holes, perhaps in a correlated fashion, and maybe varying the number of layers over time. Reason himself discusses similar criticisms (Reason, 1995, 1997); but the simpler mechanistic thinking implicit in Fig. 1 still pervades thinking in much of reliability engineering (Perneger, 2005). The model visually emphasises a reductionist approach to HRA and may thus 'wrong-foot' the users of reliability analysis methodologies leading them to miss some of the key factors and mechanisms that should be built into their models; and, perhaps, put too much trust in the combined effect of several safety barriers. For example, it could be argued that the model struggles to fully represent the motives that might accompany deliberate violations of procedure, the creeping

[2] One should not take too chronological perspective on first, second and third generation HRA methods. Some of those developed earliest did make attempts to account for contextual effects (Adhikari et al. 2008).

loss of vigilance with respect to a safety culture or the very real opportunities for the bypass of barriers in most technological systems.

We note that there is an established literature stemming from a range of work in France on the need to moderate reductionist, decomposable approaches to human reliability – or as they sometimes term it, 'human factors of reliability' – with an understanding of organisational, management and process contexts which can introduce dependencies (Fadier, 2008; Fadier and Ciccotelli, 1999; Fadier and De la Garza, 2006; Leplat, 1994).

3. Human behaviour and human error

Human behaviour is complex and often non rational. For instance, it seems sensible to use modern technological advances to make the physical components of a system safer. But there is some evidence that making subsystems safer could make the overall system less safe because of the propensity of humans to take less care personally when a system takes more care (Adams, 1988; Hollnagel, 1993). In this section we survey some recent findings from behavioural decision studies and consider how this area of theory and research can add to HRA. We do not focus on error behaviours *per se,* but take a more holistic approach. We do this for three reasons.

First, the error focus of HRA models may be too narrow (Hollnagel, 1998, 2000). Errors are just one of a range of behavioural products of a number of individual and organisational precursors; they are not a class of behaviours that are entirely distinct from other behaviours and thus should not be considered in isolation. In the organisational context, it is often an external system or judgement that categorises a behaviour as an error rather than the behaviour itself being inherently and indisputably wrong.

Second, models of HRA that explicitly include human factors typically focus on cognitive aspects of decision making. Recent developments in the modelling of decision making emphasise the dual influences of cognition and emotion on decision outcomes (French et al., 2009; Loewenstein et al., 2001; Slovic et al., 2004). The integration of emotions and cognition models of decision making has improved the ability of such models to understand and predict behaviour (Phelps, 2006). Furthermore, such an integrated approach is highly relevant to the risk-related decision making typically found within safety critical industries (Fenton-O'Creevy et al., 2008; Finucane et al., 2000).

Third, the use of high reliability systems designed and engineered to minimise errors and hazards has both benefits and disadvantages. It is of course important that systems are designed to be as safe as possible. However, the reliance on such systems can cause biases and flaws in decision making. The *risk thermostat* model suggests there is a dynamic interaction between actors' perceptions and behaviours and their environment (Adams, 1988; Wilde, 1982, 1998). People will adjust their behaviour to be more or less risky, as appropriate for their preferences and their situation, perhaps relying on one safety system to protect them from the risk of failing to operate another. A high profile example is the leak in a modern plant at Sellafield mentioned previously. There was a belief that such a modern plant could not suffer from leaks or other failures. In the context of the 'new plant' culture and other management imperatives, it was too easy to ignore inconclusive but pertinent readings and observations. It is also noteworthy that this 'new plant' culture was implicated in two previous smaller incidents at Sellafield (Adhikari et al., 2008; Board of Inquiry, 2005). Marcus and Nichols (1999) discuss similar behaviours in which warning signs were not heeded and suggest that other priorities for limited resources make it too easy to drift towards what they term the 'safety border'.

Real human judgement and decision making is not as rational and analytic as one might wish. Since the early 1980s, psychologists have distinguished between two different forms of thinking (Chaiken et al., 1989)[3]:

- System 1 thinking, often referred to as 'intuition' or 'gut reaction' that involves a superficial analysis/interpretation of the relevant information based on much simpler forms of thinking on the fringes or outside of consciousness.
- System 2 thinking, characterised by conscious analytical thought that involves a detailed evaluation of a broad range of information, often based on a rule that is assumed to provide the 'correct' answer or solution.

While formal risk assessment techniques have the characteristics of System 2 thinking, system operators may use System 1 thinking in their day-to-day operations and responses to events. For example, a nuclear power plant is the outcome of considerable complex analysis, research and design, i.e. System 2 thinking. The operators of such a plant, however, do not typically engage in the same kind of analytical thinking as the system engineers and designers. The operators' work comprises much more routine procedures and, where complex problems are faced, there is potential for operators to make them more manageable through System 1 heuristics. It has become common to refer to much of System 1 thinking as involving 'heuristics and biases', because of its deviation from the more rational, analytic System 2 thinking, though that terminology is as pejorative as the constant use of the term 'human error' in HRA which we reject in this article.

There is an extensive literature on decision making heuristics and biases (French et al., 2009; Kahneman et al., 1982; Kahneman and Tversky, 2000). Numerous studies have demonstrated the existence of systematic and robust cognitive biases, and are well summarised by Bazerman (2006). For example, emotionally-laden or otherwise individually salient information is recalled easily and likely to be considered as significant to a decision when more objective evidence shows that other types of information are more important to a decision. The processes that drive such biases have not arisen without reason – we cannot take into account all the information that surrounds us and so we need to select information to attend to in order for any action to be taken. The work of Gigerenzer and colleagues has shown that some heuristics can improve decision making by providing rapid mechanisms for recall of salient information and execution of choice behaviours (Goldstein and Gigerenzer, 2002). However, such biases can be problematic. For example, Willman et al. (2001) and Fenton-O'Creevy et al. (2003) explored the dislocation between pure financial theories and the collective and individual behaviours of market traders. Their research showed that biases led to ineffective decision making and reduced performance.

It is of concern that very little use of this extensive, often empirically based literature has been made in developing HRA methodologies. Indeed, the mechanistic approach common to many such methodologies based on fault tree representations of human action assumes that the operators are using System 2 thinking when in all probability their intuitive responses and actions are guided by System 1 thinking (Bargh et al., 1996).[4] HRA methodologies should model the thinking and behaviours that are likely to occur rather

[3] There is an unfortunate conflict of terminology here between our use of 'system' to mean the entire plant and processes which is at risk and 'systems of thinking' as referred to in the psychological literature. We use the phrases 'System 1 (or 2) *thinking*' to distinguish the latter.

[4] Of course, one might hope that if operators have been subject to many training exercises, then their responses may be closer to those that would arise from System 2 thinking.

than more rational, analytic actions and responses that one should like to think would occur.

In fairness to some current approaches to quantitative HRA, their proponents would not claim to be modelling actual behaviour, whether it be driven by System 1 or System 2 thinking; nor to be seeking a 'correct' answer to a quantitative problem. When risk analysis is used formatively, its purpose is to understand better systems and identify the key drivers of risk, rather than chase quantified estimates *per se*. Current HRA methods may help identify the key drivers relating to human behaviour, irrespective of what is going on inside people's heads and whatever organisational and environment contexts that surround them. However, such approaches do need data: and while there is generally no great problem in finding data relating to normal operations, appropriate data are – fortunately! – sparse in most contexts relating to serious system failures.

If we are to model actual behaviour in a variety of circumstances, then the concept of self-regulation may be needed. Individual self-regulation is defined as the internal and behavioural adjustments that function to maintain factors such as cognitions, emotions and performance within acceptable limits (Lord and Levy, 1994). This approach to modelling behaviour proposed that behaviour is goal orientated and there are internal, hierarchical processes that enable people to put thoughts into actions through activation and inhibition of decision making processes (Carver and Scheier, 1981). Some of the decision processes take place at a subconscious level and never reach conscious deliberation, a process called automaticity (Bargh and Chartrand, 1999). Thus, there is a dynamic interaction between people and their environment that is designed for effective behaviour. Models of decision making and behaviour that incorporate optimal levels of functioning have a long history and a range of organisational applications. For example, Yerkes and Dodson (1908) introduced an inverted U model of the association between performance and arousal. More recent models of work performance show similar patterns: some effort and pressure can be effective, too much of either leads to burnout (Schaufeli and Bakker, 2004). The organisational context must be considered both as an influence on individual level decision making and as an integral outcome of individual and group decision making processes. Choices are made at all levels of organisational design that are potentially subject to the same processes of automaticity, flawed biases and self-regulation as individual decision making.

This recognition that human behaviour is complex and driven by a range of internal and external factors leads us to question the value of terminology such as 'error', 'slip' or 'failure' within HRA. Human errors and faults are socially defined events: a perfectly reasonable action to one person may be an unreasonable failure to another (Hollnagel, 2000). Furthermore, however well judged a decision may be *a priori*, it may through 'ill fortune' lead to unwanted outcomes. Hence what may seem an error in hindsight may not be the outcome of irrational or erroneous choice. We should focus more on human *behaviour* in individual, group and organisational contexts and recognise its potential involvement in system failure – without the pejorative judgement of whether that behaviour is aberrant in any sense. For example, in the Three Mile Island Incident (Commission on the Three Mile Island Accident, 1979) the initiating event – the formation of a hydrogen bubble which forced down cooling water exposing the core – had not been anticipated in the reactor's design or safety studies. The operators not only did not recognise what was happening, but also had never anticipated that it might. It was an incident beyond their experience and imagination, in a very real sense outside of scientific and engineering knowledge as it stood then. The operators behaved entirely sensibly and in accordance with their mental models of what they believed was happening. There was no error in their behaviour in this respect, not at least in the sense of human error within HRA theory. As we build and operate more and more complex systems, we should recognise that it is inevitable that we will encounter unanticipated events and conditions. Risk and reliability analyses need to take account of human responses to these and, although those responses may indeed lead to untoward outcomes, it is far from clear that they should be dubbed errors.

4. High reliability organisations

The past 20 years has seen several studies of *high reliability organisations* (HROs), which Roberts (1990) defined as organisations failing with catastrophic consequences less than one time in 10,000. These studies recognise that certain kinds of social organisation are capable of making even inherently vulnerable technologies reliable enough for a highly demanding society.

An HRO encourages a culture and operating style which emphasises the need for reliability rather than efficiency (Weick, 1987). As organisations, HROs emphasise a culture of learning, although they clearly do not rely in any sense of learning from mistakes! Instead, HROs resort to learning from imagination, vicarious experience, stories, simulations and other symbolic representations (Weick, 1987). They emphasise a culture of sharing of learning and knowledge, of mental models: 'heedful inter-relating' (Weick and Roberts, 1993), 'collective mindfulness' (Weick et al., 1999), 'extraordinarily dense' patterns of cooperative behaviour (La Porte, 1996) and 'shared situation awareness' (Roth et al., 2006). Usually HROs apply a strategy of redundancy (Rochlin et al., 1987) with teams of operators 'watching each others backs'. As noted, it is suggested that teams share common mental models of both their internal organisational processes and the external world (Mathieu et al., 2000; Smith-Jentsch et al., 2005). Redundancy may increase complexity of operations as it makes the operations system less easily understood or opaque (Perrow, 1984; Sagan, 1993). However, redundancy also increases the probability or chance of getting adequate information to solve probable dangers, consequently reducing the risks arising from complexity rather than increasing them. When necessary, HROs try to decentralise the authority of senior teams or management responsible for decision making. Rijpma (1997) suggests that HROs use decentralisation to enable those working closest to any problems to address and solve them as they emerge or become apparent. Using this method rapid problem solving is achieved, resulting in an increase in reliability and reduction of the risk of accidents occurring in highly critical situations. This decentralisation may increase the complexity of the organisation as knowledge and lines of authority need to be distributed, but La Porte (1996) suggests the balance of these opposing effects can lie in the direction of higher reliability.

There are several challenges that have been mounted to the HRO line of work. First, some suggest that HRO perspectives are heavily functionalist and neglect politics and group interests (Perrow, 1994; Sagan, 1993, 1994). A second criticism relates to the absence of validation for the empirical studies underpinning HRO theory (Clarke, 1993; Perrow, 1994; Sagan, 1993). Critics argue that the context of some of the most important HRO studies, e.g. on the flight decks of aircraft carriers, is misleading, with evidence of safety only in simulated rather than actual operations. Others argue that the mechanisms and qualities that are said to underlie the achievement of high reliability are neither particularly characteristic of HROs nor unequivocally good for reliability. But the HRO work has given us an insight into the way in which error and failure is managed by social organisations, and how collective, rather than individual, phenomena like collective mindfulness (Weick et al., 1999) are what produce reliability in the face of supposedly unreliable individuals and unreliable technologies. The emphasis of HRA on individuals and on atomised tasks therefore misses the probability that collective actions and behaviours might lead to or avert system failure.

There would seem to be considerable potential for formative uses of HRA to influence the development of HRO theory, at least in so far as it can be applied in system and organisational design; and vice versa, complementing the work of, e.g., Grabowski and Roberts (1999).

5. Decision contexts

There is a further aspect of context that HRA should consider: decision context. The judgements and decisions needed of humans in a system can vary from those needed to perform mundane repetitive operational tasks through more complex circumstances in which information needs to be sought and evaluated to identify appropriate actions to the ability to react to and deal with unknown and unanticipated. Decision processes will vary accordingly. Design decisions can inadvertently introduce further risks to the system that arise from limitations inherent in human foresight. This means that the appropriate HRA methodology to assess the risks associated with the human decision making behaviour may vary with the details of that context.

Cynefin is a conceptual framework developed by Snowden which, among other things, offers a categorisation of decision contexts (Snowden, 2002; Snowden and Boone, 2007). The Cynefin model roughly divides decision contexts into four spaces: see Fig. 2. In the *known space*, or the Realm of Scientific Knowledge, the relationships between cause and effect are well understood. All systems and behaviours can be fully modelled. The consequences of any course of action can be predicted with near certainty. In such contexts, decision making tends to take the form of recognising patterns and responding to them with well rehearsed actions. Klein (1993) discusses such situations as recognition primed decision making. In the *knowable space*, the Realm of Scientific Inquiry, cause and effect relationships are generally understood, but for any specific decision there is a need to gather and analyse further data before the consequences of any course of action can be predicted with any certainty. Decision making can be proceduralised with clear guidance decided *a priori*. In the *complex space*, often called the Realm of Social Systems though such complexity can arise in environmental, biological and other contexts, decision making situations involve many interacting causes and effects. Knowledge is at best qualitative: there are simply too many potential interactions to disentangle particular causes and effects. Before decisions can be made, it is necessary to think widely, explore issues, frame the problem and develop broad strategies that are flexible enough to accommodate changes as the situation evolves. Much judgement and expertise will be needed in making the decision itself. Finally, in the chaotic space, situations involve events and behaviours beyond our current experience and there are no obvious candidates for cause and effect. Decision making cannot be based upon analysis because there are no concepts of how separate entities and predict their interactions. Decision makers will need to take probing actions and see what happens, until they can make some sort of sense of the situation, gradually drawing the context back into one of the other spaces. The boundaries between the four spaces should not be taken as hard. The interpretation is much softer with recognition that there are no clear cut boundaries and, say, some contexts in the knowable space may well have a minority of characteristics more appropriate to the complex space.

The Cynefin framework provides a structure in which to articulate some concerns about the use if HRA in risk and reliability analysis and in relation to hro studies.

- First generation HRA methodologies and arguably most of second and third generation ones focus on repetitive, operational tasks that lie in the known or, perhaps, knowable spaces. Yet many of the perceived risks in modern systems arise because of their inherent complexity (Perrow, 1984, 1994). In other words, we need be concerned with human behaviour as managers and operators strive to deal with events happening in the complex or even chaotic spaces. The Chernobyl Accident was initially managed as if it were in the known and knowable spaces, yet it was one of the most complex socio-technical accidents that have occurred (French and Niculae, 2005). In the Three Mile Island Accident initially there was no conceptual understanding of the processes by which a hydrogen bubble might form and hence decision making in the first hours and days of handling the incident took place in the chaotic space.
- It is informative to read HRO studies from the perspective of Cynefin. For instance, Weick's (1987) discussion moves from discussions of how air traffic controllers manage flights in a highly reliable way – a repetitive task in the known/knowable spaces – and uses these to discuss how teams might react to complex events such as Bhopal, the decision to launch Challenger and the Three Mile Island Accident. It is far from clear that organisational practices that enable repetitive, intrinsically dangerous operations to be carried out safely can be used to develop organisational preparedness dealing with complex situations that bring many risks, some quite unanticipated. (For discussions of the tension between operational risk management practice, and incident preparedness and management, see, e.g., Jalba et al., in press; Pollard et al., 2009).

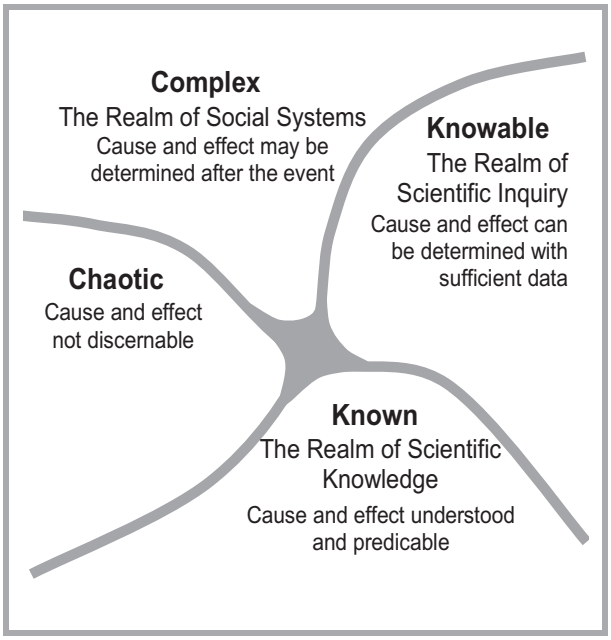

Fig. 2. Cynefin.

The appropriateness of any HRA methodology may depend on the context that is being assessed. As is the case with all risk methodologies, the characteristics of the risk and the availability of data to support the application of specific tools and techniques has a forceful influence on their feasible use. Are we considering a repetitive task that an operator performs in the normal course of events? In this case we need modelling approaches that fit with behaviours in the known domain. Or are we looking at the response of an operator to something unexpected that may herald an unanticipated departure of the system from its normal operating characteristics? In this case we need modelling behaviours for the knowable, complex or even chaotic domain. For repetitive events the key contextual pressures on operators that may modify

their behaviour are likely to relate to complacency and organisational issues such as excessive workloads or requirements to work at the same task too long. External pressures and distractions such family problems or a national sporting event are more likely to affect behaviour in repetitive normal operations than in responding to the unexpected. In responding to events ranging from an indication of departure from normal operations to a full blown crisis, adrenaline, the importance of the matter, as well as cognitive interest are likely to focus the mind. So the operators' performance is more likely to be affected by issues such as cognitive overload, miscommunication between several operations and a range of behaviours that we commonly call panic! Organisational contexts that affect the operators' responses relate to, *inter alia*, the provision of training, including emergency simulations in a variety of scenarios, and the establishment of common mental models among response teams and, more generally, of supportive team behaviours.

Our contention is that the variety of tasks that HRA is called upon to perform and the range of contexts in which it is applied are so great that it would be optimistic in the extreme to expect one methodology to be sufficient to meet these requirements. Hollnagel (1998) recognised this, though his suggestion of two methods probably does not take us much further forward, particularly as his basic method is more of a screening method for his extended approach rather than appropriate to a different set of circumstances. What we believe is needed is a portfolio of HRA methods. The characteristics of each need to be well understood so that we can determine the appropriate contexts for its application and appreciate its accuracy. It is also important to work out a way of integrating them so that we do not perpetuate the fallacy of thinking tasks can be divided up and broken down, and methods can be selected in isolation.

6. Toward an extended model of HRA

Summative HRA and related approaches emphasise quantification and prediction. While cognitive understanding of people and cultural perspectives on organisations are acknowledged, the gulf between these and quantitative risk models is generally considered too significant to be bridged. Yet the conjoining of these approaches could yield a superior model of safety critical organisations and the people working within them. In the short term, exploring the interfaces between HRA and behavioural, organisational and related studies is likely to benefit formative analyses to support the design and operation of complex systems. The barriers to the quantification needed in summative analyses are currently too substantial – we do not have sufficiently developed and validated models of behaviour and organisations to provide the precision needed. Moreover, progress in improving and developing quantitative HRA methods is likely to proceed most quickly in relation to tasks and activities falling in the known and, perhaps, knowable Cynefin spaces. Successful quantitative modelling of such tasks and activities depends on having sufficient data to develop and validate models. For systems that are long established or straightforward developments thereof, we are likely to have useful data. For novel systems we might generate such data by involving operators in simulations of component tasks and activities in known and knowable spaces. Ĉepin (2008) has suggested as much, though without the language of Cynefin. Ĉepin's modelling, as might be expected from our discussion, is focused on probabilistic assessments of errors of omission and commission. His proposed development focuses on manufactured situations with tight parameters. While additional data gathered within such a paradigm would add considerable utility to HRA models, there remains the issue of scope. The approach cannot be easily extended to tasks and activities in the other spaces. By definition, in the complex space we have neither sufficient qualitative understanding nor relevant data to develop quantitative HRA models that predict individual, group and organisational human behaviour and how these may impact overall system reliability and safety.

Thus we believe that the dominant HRA paradigm, suited as it is for the known and knowable spaces, needs to be complemented by paradigms developed specifically for the complex space. In achieving this, we will need to move away from many systems engineering approaches in which hazards are purportedly designed out of a system. Complex systems involve some human activity if only in their design and hence are susceptible to some risk arising from human behaviour. Such systems engineering approaches may work in the known or knowable spaces – and there the question is moot. Even the simplest systems in the known space need to be designed and that is a human activity. Moreover the risk homeostasis model suggests that there can be an over-reliance the safety promised by the system and a concomitant increase in overall risk. But in the complex space, we have no such hopes that current approaches can design hazards out of the system.

It will not be easy to develop models for new HRA paradigms suitable for the complex space. For instance, organisational behavioural studies can be useful in identifying the individual, cultural and organisational factors relevant to system safety; but they do not lend themselves to simple quantification; indeed, it is the very nature of the complex space that quantification is difficult if not impossible on current knowledge. We are also limited both by the availability of data from real incidents and from the generalisability of laboratory based studies. Some commentators, notably Le Coze (2005), consider the question of whether current forms of organisations can lend themselves to effective modelling. Furthermore, linear models of cause and effect cannot be simply applied (Morin, 1977). Le Coze provides a useful analysis of organisational theories, and their limitations. He proposes that approaches from complexity theory[5] (Morin, 1999; Prigogine, 1994; Simon, 1996) could assist in integrating methodologies. One of the contributions of complexity theory is the guiding philosophy that complex problems cannot be meaningfully decomposed and retain utility since the whole is greater than the sum of its parts. Le Coze concludes by emphasizing the need for holistic approaches to organisations with additional data from both organisational events and empirical studies.

Another family of approaches that might lead to a broadened conceptualisation of HRA in the complex space are the socio-technical (Mumford, 2000). For instance, Reiman and Oedewald (2007) propose that safe and effective organisations can arise only when there is integration of organisational culture and organisational activities. Their model includes a range of qualitative and quantitative methods designed to elicit descriptions of the cultural features and the organisational core tasks resulting in a thorough understanding of alternative ways to approach organisational thinking, strengths and weaknesses of practices and opportunities to create dialogue regarding the effectiveness of work. Reiman and Oedewald's paper represents a useful contribution to the development of the field of socio-technical systems and their potential links with more quantitative approaches to error and risk. What they do not encompass are individual approaches to understanding organisational behaviour. Their work represents another step in the right direction but there is a long way to go before human activities and behaviours in complex space can be modelled sufficiently for quantitative HRA.

None of the above addresses activities and behaviours which might arise if through some unanticipated event the system 'moves into' the chaotic space: e.g. the unanticipated formation

[5] There are differences between complexity theory and Snowdon's notion of the complex space in Cynefin, but there are also similarities and these link Le Coze's and our arguments.

of a hydrogen bubble in the Three Mile Island Incident (Commission on the Three Mile Island Accident, 1979; Niculae, 2005). By definition these characteristics cannot be represented in a model – certainly not in anything other than a schematic manner – simply because the chaotic space represents that part of our environment that we do not understand yet and so cannot predict.

So to take stock: current quantitative HRA methodologies seem applicable to behaviours and activities in the known and knowable spaces. There are the barest hints of how some quantitative models might be developed to predict the impacts of human activities and behaviours in the complex spaces; and, by definition, it is logically inconceivable that we can develop quantitative models for the chaotic space. Thus it is not currently possible to perform summative risk and reliability analyses for any system in which human behaviour and activity can enter the complex or chaotic spaces. Governments and regulators should be concerned because this accounts for the majority of the technological systems currently being operated and commissioned. This does not mean that they are unreliable or unsafe; only that we cannot assure their reliability or safety to within some negligibly small probability. But there are ways forward.

Firstly and most immediately, we can look to formative uses of HRA, the behavioural and organisational sciences and many other related disciplines to inform the design of organisational and management structures and the establishment of appropriate safety cultures to improve the systems that we have and are designing. This will not be easy because the imperatives that drive this approach fly in the face of the dominant reductionist thinking in risk and reliability communities. One cannot simply decompose systems into smaller subsystems, focus on these in turn and expect these to represent the total system, because culture, organisational structures and other drivers of human behaviour correlate actions, judgements and decision making in the different subsystems. Modern perspectives on risk demand a systemic rather than an atomised perspective of the technical, human and organisational features of systems. Further, because many systems have shared, and arguably, often fragmented responsibilities for management and risk management (e.g. flood defence, social care, biosecurity in the food chain), one needs to take a more holistic perspective. The conceptualisation offered by Cynefin may again give us a way forward. The simple visual categorisation of different decision contexts has proven very successful in one of the author's experiences in helping in problem formulation and issue structuring (Franco et al., 2006, 2007; Mingers and Rosenhead, 2004; Rosenhead and Mingers, 2001). The managers who decide on the choice of managerial system, its components and operational processes could map these onto a Cynefin diagram. The discussions and deliberations that would occur as they undertook this would naturally surface many issues that their design and management decisions would need to address. In other words, we propose a careful use and reflection upon a Cynefin mapping would augment current hazard identification procedures and make clearer some of the issues relating to human behaviour that management will face in operating the system. When they identify that issue although important lies in the known or knowable space they can look to current HRA – or, preferably, somewhat enhanced – models to guide their thinking and planning. But when discussion identifies an issue as lying in the complex space then they will to rely much more on judgement and put into place management processes that can deal with behaviours more subtly than seeking to police against 'slips, errors, and omissions'.

We also believe that in time it will be possible to develop better quantitative HRA methodologies to give wider assurance at the summative level. But it is unlikely that this will lead to single methodology. Rather we will need a multi-faceted approach that combines empirically validated HRA models for the known and knowable spaces with more judgementally based methods for the complex space. The Cynefin model suggests a broad framework with which to categorise the human tasks and activities in system to determine which form of HRA modelling would be most appropriate; but it is only a broad framework. To develop this methodology it will probably need extending to recognise, among other things:

- whether the human behaviours and activities take place at the individual, group, organisational level;
- the wider organisational context – including strategic and economic imperatives – in which the teams and local management structures are embedded;
- the team and local management structures which set the local context in which the operators work;
- the cultural context and – including misplaced trust in other safety barriers in the system – in which the operators find themselves;
- external influences, particularly those arising from larger external and societal pressures;
- the historical context, including perhaps the lack of recent incidents leading to a growth of complacency.

In Adhikari et al.(2008) we outline a programme of research and benchmarking that may help us develop such a multi-faceted portfolio of HRA methodologies that may eventually provide much better summative guidance on the risks inherent in complex systems.

None of this will be easy and it will only be possible if we can break the current mechanistic paradigms that permeate the risk and reliability communities. We need to move on from the Swiss Cheese model.

7. Conclusion: a message for managers

The key point that we have been trying to convey in this paper is the current dislocation between the mechanistic reductionist assumptions on which current HRA methodologies are primarily built and our current understandings of human and organisational behaviour. We must bring these into better register. Managers, regulators, politicians and the public need to beware of this lest they believe the numbers that are sometimes touted about the safety of our systems. This should not be read as a manifesto for Luddism. We are not against the development of more and more complex systems, providing that they bring benefits, of course. Nor are we against risk *per se*. Rather we are concerned at the prevalence of overconfidence in our ability to assess the risks that arise from human behaviour. We need to take the numbers with that 'pinch of salt', recognising that when we build complex systems our uncertainty is greater than the raw numbers suggest and we need to monitor and watch for the unanticipated. As is often the case with the application of risk and reliability tools, the valuable insight comes from a systemic and often qualitative understanding of which systems features 'drive' the risk, rather than from the risk estimates *per se*.

We in the research community have much to do. But so does the management community. It is too easy to trust the assurances of current risk and reliability analyses which promise that the chance of an untoward event is small, to believe in the cumulative effect of 'independent' safety barriers and to manage the subsystems separately unaware of the interconnections between them that organisational culture and human behaviour bring. Human reliability has too long been treated as something that relates to individuals. It needs to be seen and managed at the organisational level. The key question is not how likely is an individual's behaviour is to impact a system, but how well the organisational structures around

and within that system enable the system to run safely and reliability, and how well they will recover if an untoward event threatens or happens.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (Contract number: EP/E017800/1). We are grateful to our co-investigators and colleagues on this: Sondipon Adhikari, Clare Bayley, Jerry Busby, Andrew Cliffe, Geeta Devgun, Moetaz Eid, Ritesh Keshvala, David Tracy and Shaomin Wu. We are also grateful for many helpful discussions with Ronald Boring, Roger Cooke and John Maule.

References

Adams, J., 1988. Risk homeostasis and the purpose of safety regulation. Ergonomics 31 (4), 407–428.
Adhikari, S., Bayley, C., Bedford, T., Busby, J.S., Cliffe, A., Devgun, G., Eid, M., French, S., Keshvala, R., Pollard, S., Soane, E., Tracy, D., Wu, S., 2008. Human Reliability Analysis: A Review and Critique. Manchester Business School, Booth Street West, Manchester M15 6PB.
Aven, T., 2003. Foundation of Risk Analysis: A Knowledge and Decision Oriented Perspective. John Wiley and Sons, Chichester.
Bargh, J.A., Chartrand, T.L., 1999. The unbearable automaticity of being. American Psychologist 54, 462–479.
Bargh, J.A., Chen, M., Burrows, L., 1996. Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology 71, 230–244.
Barlow, R.E., Proschan, F., 1975. Statistical Theory of Reliability and Life Testing. Holt, Reinhart and Winston, New York.
Barriere, M., Bley, D., Cooper, S., Forester, J., Kolaczkowski, A., Luckas, W., Parry, G., Ramey-Smith, A., Thompson, C., Whitehead, D., Wreathall, J., 2000. NUREG-1624: Technical Basis and Implementation Guidelines for a Technique for Human Event Analysis (ATHEANA). US Nuclear Regulatory Commission.
Bazerman, M.H., 1999. Reviews on decision making. Administrative Science Quarterly 44 (1), 176–180.
Bazerman, M.H., 2006. Managerial Decision Making, sixth ed. John Wiley and Sons, New York.
Bedford, T., Cooke, R., 2001. Probabilistic Risk Analysis: Foundations and Methods. Cambridge University Press, Cambridge.
Board of Inquiry, 2005. Fractured Pipe with Loss of Primary Containment in the THORP Feed Clarification Cell. British Nuclear Fuels Limited.
Boring, R.L., 2007. Dynamic Human Reliability Analysis: Benefits and Challenges of Simulating Human Performance European Safety and Reliability Conference (ESREL 2007). INL/CON-07-12773, Idaho National Laboratory.
Carver, C.S., Scheier, M.F., 1981. Attention and Self-Regulation: A Control Theory Approach to Human Behavior. Springer Verlag, New York.
Ĉepin, M., 2008. Importance of human contribution within the human reliability analysis (IJS-HRA). Journal of Loss Prevention in the Process Industries 21 (3), 268–276.
Chaiken, S., Liberman, A., Eagly, A.H., 1989. Heuristic and systematic information processing within and beyond the persuasion context. In: Uleman, J.S., Bargh, J.A. (Eds.), Unintended Thought. Guilford, New York, pp. 212–252.
Clarke, L., 1993. Drs Pangloss and Strangelove meet organizational theory: high reliability organizations and nuclear weapons accidents. Sociological Forum 8, 675–689.
Commission on the Three Mile Island Accident, 1979. Report of the President's Commission on the Accident at Three Miles Island. US GPO, Washington, DC.
Courtois, P.-J., Littlewood, B., Strigini, L., Wright, D., Fenton, N., Neil, M., 2000. Bayesian belief networks for safety assessment of computer-based systems. In: Gelenbe, E. (Ed.), System Performance Evaluation: Methodologies and Applications. CRC Press, pp. 349–363.
Fadier, E., 2008. Editorial of the special issue: design process and human factors integration. cognition. Technology and Work 10 (1), 1–5.
Fadier, E., Ciccotelli, J., 1999. How to integrate safety in design: methods and models. Human Factors and Ergonomics in Manufacturing & Service Industries 9 (4), 367–379.
Fadier, E., De la Garza, C., 2006. Safety design: towards a new philosophy. Safety Science 44, 55–73.
Fenton-O'Creevy, M., Nicholson, N., Soane, E., Willman, P., 2003. Trading on illusions: unrealistic perceptions of control and trading performance. Journal of Occupational and Organizational Psychology 76 (1), 53–68.
Fenton-O'Creevy, M., Soane, E., Nicholson, N., Willman, P., 2008. Thinking, feeling and deciding: the influence of emotions on the decision making and performance of traders. In: Academy of Management Conference, Anaheim, California.
Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M., 2000. The affect heuristic in judgments of risks and benefits. Journal of Behavioral Decision Making 13, 1–17.
Forester, J.A., Kolaczkowski, A., Lois, E., Kelly, D., 2006. NUREG-1842: Evaluation of Human Reliability Analysis Methods against Good Practices. US Nuclear Regulatory Commission, Washington, DC.
Franco, A., Shaw, D., Westcombe, M., 2006. Problem structuring methods I. Journal of the Operational Research Society, 757–878.
Franco, A., Shaw, D., Westcombe, M., 2007. Problem structuring methods II. Journal of the Operational Research Society, 545–682.
French, S., Maule, A.J., Papamichail, K.N., 2009. Decision Behaviour, Analysis and Support. Cambridge University Press, Cambridge.
French, S., Niculae, C., 2005. Believe in the model: mishandle the emergency. Journal of Homeland Security and Emergency Management 2 (1).
Goldstein, D.G., Gigerenzer, G., 2002. Models of ecological rationality: the recognition heuristic. Psychological Review 109 (1), 75–90.
Grabowski, M., Roberts, K.H., 1999. Risk mitigation in virtual organizations. Organization Science 10, 704–721.
Hannaman, G.W., Spurgin, A.J., Lukic, Y.D., 1984. Human Cognitive Reliability Model for PRA Analysis. Draft Report NUS-4531, EPRI Project RP2170-3. Electric Power and Research Institute, Palo Alto, CA.
Helmreich, R.L., 2000. On error management: lessons from aviation. British Medical Journal 320 (7237), 781–785.
Hollnagel, E., 1993. Human Reliability Analysis: Context and Control. Academic Press, London.
Hollnagel, E., 1998. Cognitive Reliability and Error Analysis Method – CREAM. Elsevier Science, Oxford.
Hollnagel, E., 2000. Looking for errors of omission and commission or the hunting of the Snark revisited. Reliability Engineering and System Safety 68, 135–145.
Høyland, A., Rausand, M., 1994. System Reliability Theory. John Wiley and Sons, New York.
Hrudey, S.E., Hrudey, E.J., Charrois, J.W.A., Pollard, S.J.T., 2006. A 'Swiss Cheese' Model Analysis of the Risk Management Failures in the Fatal Walkerton Outbreak. IWA World Water Congress and Exhibition, Beijing, China.
International Atomic Energy Agency, 1991. The International Chernobyl Project: Technical Report. IAEA, Vienna.
Jalba, D., Cromar, N., Pollard, S., Charrois, J.W.A., Bradshaw, R., Hrudey, E., in press. Safe drinking water: critical components of effective inter-agency relationships. Environment International. doi:10.1016/j.envint.2009.1009.1007.
Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.
Kahneman, D., Tversky, A. (Eds.), 2000. Choices, Values and Frames. Cambridge University Press, Cambridge.
Kariuki, S.G., Lowe, K., 2007. Integrating human factors into process analysis. Reliability Engineering and System Safety 92, 1764–1773.
Kirwan, B., 1994. Practical Guide to Human Reliability Assessment. Taylor and Francis/CRC Press, London.
Klein, G., 1993. A recognition primed decision model (RPM) of rapid decision making. In: Klein, G., Orasanu, J., Calderwood, R. (Eds.), Decision Making in Action: Models and Method. Ablex.
La Porte, T.R., 1996. High reliability organizations: unlikely, demanding and at risk. Journal of Contingencies and Crisis Management 4, 60–71.
Le Coze, J.-C., 2005. Are organisations too complex to be integrated in technical risk assessment and current safety auditing? Safety Science 43 (8), 613–638.
Leplat, J., 1994. Collective dimensions of reliability: some lines of research. European Work and Organizational Psychologist 4 (3), 271–295.
Lewis, E.E., 1994. Introduction to Reliability Engineering. Johne Wiley and Sons, Chichester.
Loewenstein, G., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. Psychological Bulletin 127 (2), 267–286.
Lord, R.G., Levy, P.E., 1994. Moving from cognition to action – a control theory perspective. Applied Psychology – An International Review (Psychologie appliquee – Revue Internationale) 43 (3), 335–398.
Lyu, M.R., 2005. Handbook of Software Reliability Engineering. IEEE Computer Society Press and McGraw-Hill Publishing Company.
Marcus, A., Nichols, M.L., 1999. On the edge: heeding the warnings of unusual events. Organizational Science 10 (4), 482–499.
Marples, D.R., 1997. Nuclear power in the former USSR: historical and contemporary perspectives. In: Marples, D.R., Young, M.J. (Eds.), Nuclear Energy and Security in the Former Soviet Union. Westview Press.
Mathieu, J.E., Heffner, T.S., Goodwin, G.F., Salas, E., Cannon-Bowers, J.A., 2000. The influence of shared mental models on team process and performance. Journal of Applied Psychology 85 (2), 273–283.
Melnick, E.L., Everitt, B.S. (Eds.), 2008. Encyclopedia of Quantitative Risk Analysis and Assessment. John Wiley and Sons, Chichester.
Mingers, J., Rosenhead, J., 2004. Problem structuring methods in action. European Journal of Operational Research 152, 530–554.
Morin, E., 1977. La method – tome I, La nature de la nature. Ed du seuil (coll point), Paris (The Method – vol I, the Nature of Nature). Seuil, Paris.
Morin, E., Lemoigne, J.L., 1999. L'intelligence de la complexité. L'Harmattan (The Intelligence of Complexity). <http://www.editions-harmattan.fr/index.asp?navig=catalogue&obj=livre&no=9175>.
Mosleh, A., Chang, Y.H., 2004. Model-based human reliability analysis: prospects and reliability. Reliability Engineering and System Safety 83, 241–253.
Mumford, E., 2000. A socio-technical approach to systems design. Requirements Engineering 5, 125–133.
Niculae, C., 2005. A Socio-technical Perspective on the Use of RODOS in Nuclear Emergency Management. The University of Manchester.

Perneger, T.V., 2005. The Swiss cheese model of safety incidents: are their holes in the metaphor. BMC Health Services Research 5, 71–77.
Perrow, C., 1984. Normal Accidents: Living with High-risk Technologies. Basic Books, New York.
Perrow, C., 1994. The limits of safety: the enhancement of a theory of accidents. Journal of Contingencies and Crisis Management 2, 212–220.
Phelps, E.A., 2006. Emotion and cognition: insights from studies of the human amygdala. Annual Review of Psychology 57, 27–53.
Pollard, S., Bradshaw, R., Tranfield, D., Charrois, J.W.A., Cromar, N., Jalba, D., 2009. Developing a Risk Management Culture—'Mindfulness' in the International Water Utility Sector (Report TC3184). Water Research Foundation, Denver, CO.
Prigogine, I., 1994. Les lois de chaos. Flammarion (The Laws of Chaos).
Reason, J., 1990a. The contribution of latent human failures to the breakdown of complex systems. Philosophical Transactions of the Royal Society of London B327 (1241), 475–484.
Reason, J., 1990b. Human error: models and management. British Medical Journal 320 (7237), 768–770.
Reason, J., 1995. Understanding adverse events: human factors. Quality Health Care 4, 80–89.
Reason, J., 1997. Managing the Risks of Organisational Accidents. Ashgate, Aldershot, UK.
Reiman, T., Oedewald, P., 2007. Assessment of complex socio-technical systems – theoretical issues concerning the use of organisational culture and organisational core task concepts. Safety Science 45, 745–768.
Ren, J., Jenkinson, I., Wang, J., Xu, D.L., Yang, J.B., 2008. A methodology to model causal relationships in offshore safety assessment focusing on human and organisational factors. Journal of Safety Research 39, 87–100.
Rijpma, J.A., 1997. Complexity, tight-coupling and reliability: connecting normal accidents theory and high reliability theory. Journal of Contingencies and Crisis Management 5 (1).
Roberts, K.H., 1990. Some characteristics of one type of high reliability organisation. Organization Science 1 (2), 160–176.
Rochlin, G.I., La Porte, T.R., Roberts, K.H., 1987. The self-designing high reliability organization: aircraft carrier operations at sea. Naval War College Review 40, 76–90.
Rosenhead, J., Mingers, J. (Eds.), 2001. Rational Analysis for a Problematic World Revisited. John Wiley and Sons, Chichester.
Roth, E.M., Multer, J., Raslear, T., 2006. Shared situation awareness as a contributor to high reliability performance in railroad operations. Organization Studies 27, 967–987.
Sagan, S.D., 1993. The Limits of Safety: Organizations, Accidents, and Nuclear Weapons. Princeton University Press, Princeton, NJ.
Sagan, S.D., 1994. Toward a political theory of organizational reliability. Journal of Contingencies and Crisis Management 2, 228–240.
Sagan, S.D., 2004. The problem of redundancy problem [sic]: why more nuclear security forces may produce less nuclear security. Risk Analysis 24, 935–946.
Schaufeli, W.B., Bakker, A.B., 2004. Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. Journal of Organisational Behavior 25, 293–315.
Simon, H., 1996. The Sciences of the Artificial. MIT Press.
Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2004. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk and rationality. Risk Analysis 24 (2), 311–322.
Smith-Jentsch, K.A., Mathieu, J.E., Kraiger, K., 2005. Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting. Journal of Applied Psychology 90 (3), 523–525.
Snowden, D., 2002. Complex acts of knowing – paradox and descriptive self-awareness. Journal of Knowledge Management 6, 100–111.
Snowden, D., Boone, M., 2007. A leader's framework for decision making. Harvard Business Review, 68–76.
Swain, A.D., Guttmann, H.E., 1983. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. NUREG/CR-1278, USNRC.
United States Nuclear Regulatory Commission, 1975. Reactor Safety Study: An Assessment of the Accident Risks in US Commercial Nuclear Power Plants.
United States Nuclear Regulatory Commission, 2002. Review of Findings for Human Performance Contribution to Risk in Operating Events (NUREG/CR-6753). US GPO, Washington, DC.
Weick, K.E., 1987. Organisational culture as a source of high reliability. California Management Review 29, 112–127.
Weick, K.E., Roberts, K.H., 1993. Collective mind in organizations: heedful interrelating on flight decks. Administrative Science Quarterly 38, 357–381.
Weick, K.E., Sutcliffe, K.M., Obstfield, D., 1999. Organizing for high reliability: processes of collective mindfulness. Research in Organizational Behavior 21, 81–123.
Wilde, G.J.S., 1982. The theory of risk homeostasis: implications for safety and health. Risk Analysis 2, 209–225.
Wilde, G.J.S., 1998. Risk homeostasis theory: an overview. Injury Prevention 4, 89–91.
Williams, J.C., 1985. HEART – a proposed method for achieving high reliability in process operation by means of human factors engineering technology. In: Proceedings of a Symposium on the Achievement of Reliability in Operating Plant, Safety and Reliability Society, NEC, Birmingham.
Willman, P., Fenton-OCreevy, M., Nicholson, N., Soane, E., 2001. Knowing the risks: theory and practice in financial market trading. Human Relations 54 (1), 887–910.
Wu, S., Hrudey, S.E., French, S., Bedford, T., Soane, E., Pollard, S.J.T., 2009. Human reliability analysis has a role in preventing drinking water incidents. Water Research 43, 3227–3238.
Yerkes, R.M., Dodson, J.D., 1908. The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurological Psychology 18, 459–482.
Zhang, X., Pham, H., 2000. An analysis of factors affecting software reliability. Journal of Systems and Software 50 (1), 43–56.

USING SPECIFIC LANGUAGE TO DESCRIBE RISK AND PROBABILITY

ANTHONY G. PATT [1] and DANIEL P. SCHRAG [2]

[1] *Boston University and Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany*
E-mail: apatt@bu.edu
[2] *Laboratory for Geochemical Oceanography, Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, U.S.A.*

Abstract. Good assessment of environmental issues, such as climate change, requires effective communication of the degree of uncertainty associated with numerous possible outcomes. One strategy that accomplishes this, while responding to people's difficulty understanding numeric probability estimates, is the use of specific language to describe probability ranges. This is the strategy adopted by the Intergovernmental Panel on Climate Change in their Third Assessment Report. There is a problem with this strategy, however, in that it uses words differently from the way lay readers of the assessment typically do. An experiment conducted with undergraduate science students confirms this. The IPCC strategy could result in miscommunication, leading readers to under-estimate the probability of high-magnitude possible outcomes.

1. Introduction

The potential impacts of climate change vary not only according to their timing and magnitude, but also according to the probability with which they will occur. Some of the most consequential potential impacts – such as rapid sea level rise due to the disintegration of the West Antarctic Ice Sheet – thankfully will probably not occur. Effective assessment of climate change allows policy-makers to take into account scientific knowledge about not only the most likely outcomes of environmental change, but also these less likely, but more consequential possibilities. A significant challenge confronting the Intergovernmental Panel on Climate Change (IPCC) and other assessment panels is to communicate the broad range of beliefs, and the uncertainties associated with those beliefs, about the future course of global climate, so that policy-makers can make responsible decisions about societal actions.

The task of communicating uncertainty is made difficult both by the disagreements within the scientific community about what the probabilities are, and by lay people's general difficulty thinking in probabilistic terms. Assessment authors must first resolve among themselves the uncertainty over uncertainty: what the probability of an event's occurring actually is when there is disagreement over that probability. Then, they must figure out how to communicate that uncertainty to a

Climatic Change **61:** 17–30, 2003.

lay audience – policy makers and the public – so that the assessment audience will be able to make effective tradeoffs with society's scarce resources.

The latest report from the IPCC, *Climate Change 2001*, systematically communicates probability using well-defined descriptive language, words such as *very unlikely* (Houghton et al., 2002). Doing so avoids having to arrive at a single point estimate for the probability of an event, or even a precise range of estimates. It also responds to the public's difficulty interpreting quantified probabilities. The IPCC strategy achieves several important objectives, such as promoting internal consensus among chapter authors and conveying a sense of confidence in outcomes of climate. At the same time, the IPCC's strategy does not exactly match people's common use of language, in which the words used to describe the probability of an event also depend on the event's potential magnitude; the IPCC is communicating probability using language commonly used to describe risk, the combination of probability and consequence.

In this paper we examine the potential biases that could result from the possible mismatch between the IPCC's use of words describing probability and people's intuitive understanding of their meaning. After background sections on people's cognitive biases interpreting probability, and the ways that assessments have commonly addressed these biases, we present the results of a simple experiment testing the use and interpretation of descriptive words to describe potential weather events. What we find is a reassuring symmetry in how people use language to describe possible events. Risk communicators exaggerate the likelihood of high consequence events, at the same time that their audience expects such exaggeration, and de-codes accordingly. The IPCC strategy, however, removes the possibility of exaggeration on the part of the communicators, since each descriptive word is assigned a specific probability range that is insensitive to event magnitude. Unless the audience adjusts – ceasing the practice of correcting for expected exaggeration – the result could be a biased under-response to high magnitude events.

2. Probability Interpretation

Both psychologists and behavioral economists have shown that people's descriptions and understanding of probabilities depend on contextual factors such as objective probability, base-rate, and event magnitude (Weber, 1994). In terms of objective probability, Kahneman and Tversky (1979) identify a weighting function people use to interpret evidence of probabilities, shown in Figure 1. People tend to overestimate the probability of relatively infrequent events (such as dying from botulism) and underestimate the probability of relatively frequent events (such as dying from heart disease). The change in people's reactions when an event's assessed probability goes from 0% to 1% is much greater than when it goes from 36% to 37% (Patt and Zeckhauser, 2002). For very small probabilities, people's responses are more binary than continuous (Kammen et al., 1997; Covello, 1990).

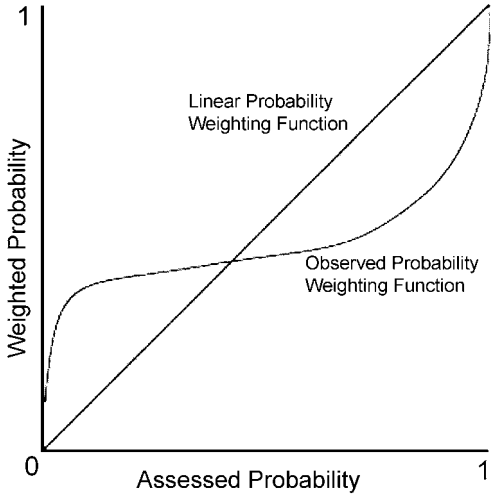

Figure 1. Probability weighting.

Below a certain threshold of concern people view the event as impossible; above the threshold, they take measures to prevent it, measures that may not be justified by the event's small probability. People are relatively insensitive to changes in assessed probability in the middle of the scale, treating all such probabilities as roughly fifty-fifty.

In terms of base rates, Wallsten et al. (1986) observe that people's interpretation of probability descriptors depends on the background frequency of an event. Hence, people interpret a 'slight chance' of rain in London as meaning a higher numeric probability than a 'slight chance' of rain in Madrid. Windschitl and Weber (1999) observe a similar phenomenon even when people are given numeric estimates of event probabilities. In one experiment, subjects are told that a person has a 30% chance of contracting a mild form of malaria during a trip to a tropical destination. Some of the subjects are led to believe that the destination is Calcutta, while others are told Honolulu. Subjects then describe, on a verbal scale, the likelihood of malaria. Those people who are told that the trip is to Calcutta tend to describe the likelihood of malaria with more certain language (choosing terms such as 'somewhat unlikely') than do the people who are told the trip is to Honolulu (choosing terms such as 'quite unlikely'). Later, the same subjects are asked to recall the numeric probability of contracting malaria. Those for whom Calcutta was the destination remember higher numeric probabilities.

In terms of event magnitude, Weber and Hilton (1990) observe people's probability word interpretation responding not only to base rates, but also to the negative utility associated with different events. In one experiment, subjects were asked to decide on the numeric probability they believe their doctor had in mind when describing the likelihood of medical conditions such as warts, stomach ulcers, and skin cancer. For each medical condition the doctor used the same probability

words, such as 'slight chance'. People's initial estimates of numeric probability are initially lower for the more serious events, such as cancer. The researchers attribute this to the base-rate phenomenon: base rates and severity are usually inversely correlated, and people generally assume higher magnitude negative outcomes are less likely. Later in the experiment, however, people were informed that the base rates were the same for the different conditions. With this new information, people show a non-linear response to event severity. As the severity of events increased, people first showed higher numeric estimates of probability. However, as the events started to become life threatening, subjects' estimates of probability begin to decrease. Hence for serious events, such as cancer, subjects again 'de-coded' the physician's language to assign a lower numeric probability than for the events of intermediate magnitude.

The sensitivities to changes in assessed probability, base rates, and event magnitude all create challenges for assessors. For example, risk communicators may have to work very hard to convince people that it is more worthwhile reducing one risk from 45% to 30% than another risk from 0.01% to 0.005%. They may have to convince people that even though a given risk has a 0% base rate – it has never happened before – it is still possible that it will happen in the future. And they will need to help people distinguish between event magnitude and probability, so that they can properly compare different risks to make more accurate decisions.

3. Uncertainty Assessment

Fortunately, scientific assessors have increasingly appeared sensitive to audience perceptions, revealed in the variety of ways they have communicated uncertainty. Some assessments fail to report highly uncertain information, or else avoid quantification of uncertainty by giving ranges of expected outcomes without clarifying the probability bounds for that range. This approach offers information that is easy to understand, yet at the same time incomplete. Patt (1999) examines the assessment of a highly unlikely yet highly consequential result of climate change – the rapid collapse of the West Antarctic Ice Sheet – across different types of assessment. He finds that the large, consensus-oriented assessments, such as the IPCC, were less likely to provide information on the event. Smaller assessments, both those conducted by advocacy groups and those responding to specific questions of their intended audience, tended to provide greater detail on the issue. There are several explanations. First, consensus within the assessment team might be difficult to achieve for high-consequence low-probability events. For example, Morgan and Keith (1995) obtained subjective probability judgments from a number of climate change scientists, using a variety of expert elicitation techniques. What they observed was disagreement, often between disciplines, with many experts' ranges failing to overlap. As events become more and more speculative, it is likely that expert opinion will diverge even more. Patt also concluded that for these extreme

events, it is possible that assessment authors would be tempted to view any treatment as counterproductive. Because people's responses to low-probability events are likely to be binary and polarized, discussion of such events may in fact lead to greater conflict within the policy community. If assessment authors see their task as building consensus, not only among themselves but also among decision-makers, then they will limit their discussion to events that are either certain or of middle-probability.

Van der Sluijs (1997), likewise, examines how the IPCC has described the range of future temperature changes associated with climate change. He observes that the range has remained fairly constant, even as new evidence has become available. Assessors were reluctant to depart from a previously stated position, and 'anchored' on the old estimate absent a compelling reason to change it. To maintain intellectual honesty, they failed to quantify the probabilities associated with that temperature range. As long as it remained unclear what a given temperature range actually meant, they could continue to use it. Like the strategy of omitting treatment of extreme events altogether, the anchoring phenomenon is a way of avoiding the rigorous treatment of uncertainty, when being rigorous could make consensus difficult, or could confuse the audience.

Other assessments – assessments of health and technological risks in particular – present quantified probability estimates. This approach offers more information but may be difficult to interpret by an untrained audience. The history of these difficulties is well documented. Leiss (1996), for example, describes three stages in risk communication practice. In the first stage, risk communicators believed that if they simply communicated their best estimates, people would use that information to make consistent tradeoffs. This strategy lasted until the 1980s, by which point it became clear that people were systematically over-reacting to some kinds of risk, and under-reacting to others. In response, risk communicators saw their jobs evolving to include more salesmanship – they would convince people of which risks were worthwhile, and which risks were not – in which the communicator was deliberately trying to bring about a specific behavior pattern that might not have occurred otherwise. Alternatively, many risk assessors and communicators started to suggest that decision-making on such issues be insulated from popular opinion (Breyer, 1993). In many cases, however, such strategies led to increased public resentment of the risk assessors and decision-makers (Freudenberg, 1996; Irwin and Wynne, 1996). The third stage, as Leiss and others (e.g., Fischhoff, 1996) see it, is characterized by a greater attention to public participation, to building partnership between risk assessors and decision-makers in developing appropriate responses to the information. The approach seems to work across issues and cultures to increase the credibility and salience of the information, and to help people respond wisely (Patt and Gwata, 2002).

Many of these considerations entered into the design consideration for the IPCC Third Assessment Report. The challenge was to provide understandable and complete information about uncertainty in a context – the written document – where

Table I
IPCC qualitative descriptors

Probability range	Descriptive term
< 1%	Extremely unlikely
1–10%	Very unlikely
10–33%	Unlikely
33–66%	Medium likelihood
66–90%	Likely
90–99%	Very likely
> 99%	Virtually certain

the audience would be unable to participate. Moss and Schneider (2000) reported to the IPCC lead authors on the communication of uncertainty, recommending a seven-step approach for describing each uncertainty. They suggested, for example, that authors should identify and describe the sources of uncertainty, document the ranges and distribution for each uncertain variable, identify the level of precision possible for describing the variable, and place the expert judgments within a formal decision-analytic framework. The IPCC authors accepted some of Moss and Schneider's recommendations, and not others. Of particular note, however, was the decision by lead authors to use specific qualitative language – words such as *likely*, *very likely*, and *virtually certain*, to describe quantitative probability ranges. Early in the report they define the probability ranges for seven qualitative descriptive terms, and then use those terms rather than numbers (see Table I). This is a more simple strategy than the one that Moss and Schneider (2000) suggest.

There may be good reasons for this approach. First, using language such as *very likely* or *virtually certain* to describe an uncertain outcome avoids the problem of experts having to reach consensus on a particular probability estimate or range. Since it may well be impossible for experts to reach consensus, the alternative to the use of such language may well be complete omission of the uncertain outcome. Obviously, it is better to describe an event than to omit it, even if the probability range is wide and not completely precise. Second, many people understand, or feel they understand, the meanings of such words better than they do accurate numbers or ranges (Wallsten et al., 1986). This is especially true for forecasts of one-time events (e.g., the chances of one meter sea level rise), as opposed to forecasts of frequent outcomes (e.g., the chances of any one person contracting malaria during a visit to Honolulu) (Pinker, 1997). To a lay audience, a numeric probability for the frequent event makes sense; the typical person stands an $X\%$ chance of contracting malaria, since X people in 100 actually do contract the disease. But for the one time event, for which there is no past data, the meaning of the $X\%$ is somewhat different. The probability estimate conveys a degree of confidence in the outcome

occurring, rather than a description of past data. The use of probability language to describe degrees of confidence, rather than numeric estimates, makes more sense to most people (Moss and Schneider, 2000). Additional information, the accurate numerical data, may simply upset this simple approach toward communicating uncertainty.

An important component of this approach, in addition to the use of words rather than numbers, is the adoption of a context-*in*dependent scale. Thus, the language the IPCC authors use to describe uncertainty depends only on the probability of the outcome, or the confidence with which they believe it will occur, and not on other characteristics of the event, such as its magnitude. However, the language that people use to discuss uncertainty and the meanings they give to various descriptors depend on the event being described and the context within which it falls: the total risk of an event. When both the communicators and the audience are using uncertainty descriptors to describe risk, and not simply probability, accurate understanding will pass from communicator to audience without bias (Brun and Teigen, 1988). But when the communicators use words to describe probabilities, and the audience still interprets them as describing risk, miscommunication can result. The result of that miscommunication could be for the audience systematically to underweight both the probability and the riskiness of high magnitude events.

4. Experiment

To illustrate how the use of context independent descriptors could be important, we conducted a simple experiment, in which we polled 152 undergraduate science students at Boston University, randomly distributing equal numbers of four different survey questions. The surveys differed across two dimensions, allowing for a controlled experiment. Half of the surveys asked subjects to translate, in the role of risk communicators, numeric probabilities into words – choosing one of the IPCC's seven descriptive terms, from *virtually certain* to *extremely unlikely* – to describe an event of 10% probability. The other half of the surveys asked subjects to assign a probability range – again one the IPCC's seven ranges, from *greater than 99% chance* to *less than 1% chance* – to an event described as 'unlikely, perhaps very unlikely'. This task is equivalent to that of an IPCC audience, making an estimate of the likelihood of an event based on the probability description they hear or read. Within each group, half the surveys asked subjects to describe or interpret the likelihood of a high-impact outcome: a hurricane due to hit land near Boston. The other half involved a low-impact outcome, early season snow flurries. Table II shows the four survey versions.

Subjects were aware that we had distributed several versions of the survey, but were not aware of how the versions differed, or the purpose of the experiment. They were also not generally aware of the IPCC's choice of language to describe uncertainty in Working Group I of the Third Assessment Report. Clearly, undergraduate

Table II
Survey versions

Communicators		Audience	
High magnitude outcome	Low magnitude outcome	High magnitude outcome	Low magnitude outcome
Imagine that you are the weather person for a Boston television station. The date is September 8, 2001.		Imagine that the date is September 8, 2001, and you are watching the weather report on television.	
You are somewhat concerned about a very powerful hurricane currently near Bermuda. Usually these hurricanes hit land in the Carolinas, or else track out to sea, but in this case conditions make it possible that the hurricane could hit land near Boston, devastating the region with sustained winds of over 100 mph and extensive flooding	You are somewhat concerned about a cold front currently over western New York State. Usually at this time of the year these fronts bring isolated thunderstorms and chilly temperatures (40s to 50s) to the region, but in this case conditions make it possible that Boston will see some snow flurries and temperatures dipping into the high 30s.	The weather person is talking about a very powerful hurricane currently near Bermuda. Usually these hurricanes hit land in the Carolinas, or else track out to sea, but in this case conditions make it possible that the hurricane could hit land near Boston, devastating the region with sustained winds of over 100 mph and extensive flooding.	The weather person is talking about a cold front currently over western New York State. Usually at this time of the year these fronts bring isolated thunderstorms and chilly temperatures (40s to 50s) to the region, but in this case conditions make it possible that Boston will see some snow flurries and temperatures dipping into the high 30s.
The National Weather Service is currently predicting the chances of this happening at 10%, and you believe this to be a good estimate. Which of the following language would you use to describe to your viewers the chances of this happening?		The weather person, whom you trust, is saying that it is unlikely, perhaps very unlikely, that this will actually happen. Based on this forecast, what do you think the chances of this event happening actually are?	
a. Extremely unlikely b. Very unlikely c. Unlikely d. Medium likelihood e. Likely f. Very likely g. Virtually certain		a. < 1% b. 1–10% c. 10–33% d. 33–66% e. 66–90% f. 90–99% g. > 99%	

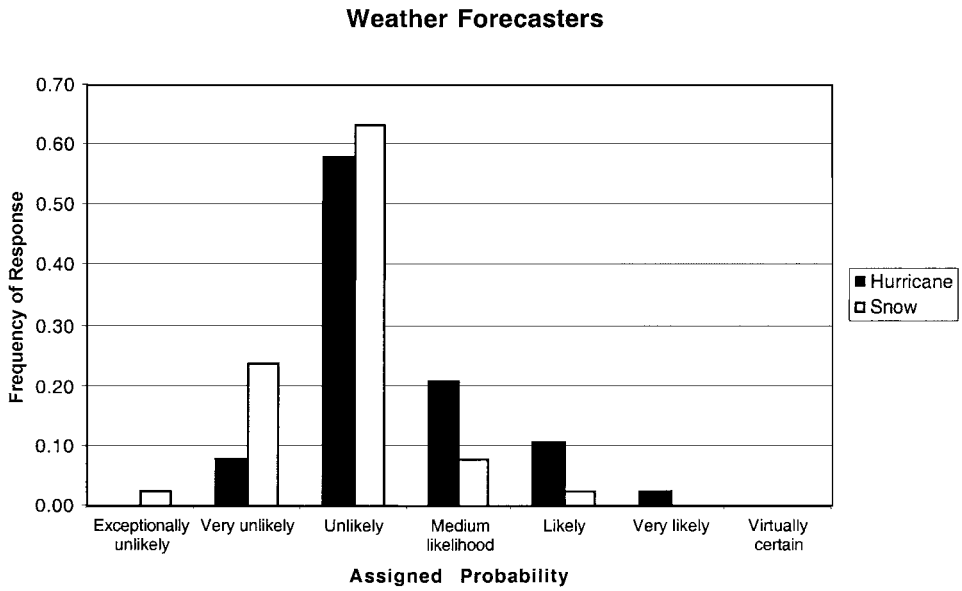

Figure 2. Communicators' probability words.

college students differ in their technical expertise from policy-makers and other readers of the IPCC report. However, what we are testing is whether there exists a basic behavioral tendency for people in general to interpret probability language describing weather events in a way that responds to event magnitude, as others have observed in the literature. It may well be that highly-trained individuals will demonstrate less of a bias. But by using college students as subjects, we can draw conclusions about people's underlying decision-making biases.

The results show significant (χ^2 test, $p < 0.01$) differences between the two outcomes across the two groups of subjects. Among communicators, subjects were more likely to use greater likelihood descriptors to describe the hurricane than to describe the snow flurries, as seen in Figure 2. While the mode descriptor for both events was *unlikely*, more subjects chose the descriptors *medium likelihood*, *likely*, and *very likely* to describe the hurricane than to describe the snow flurries; likewise, more subjects choose the descriptors *very unlikely* and *exceptionally unlikely* to describe the snowfall. Among the audience, subjects estimated lower probabilities of occurrence for the hurricane than for the snow flurries, as seen in Figure 3. The mode estimate for the hurricane was *1–10% chance*, with several subjects estimating *<1% chance*. For the snow flurries, the mode estimate was *10–33% chance*, with more subjects estimating *66–90% chance* for the snow flurries than for the hurricane.

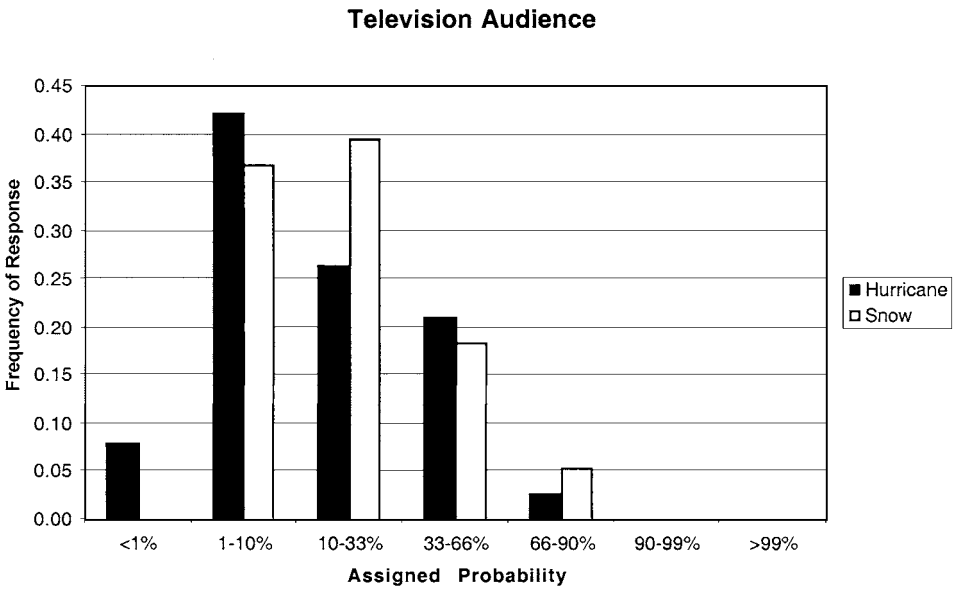

Figure 3. Audience' probability estimates.

5. Discussion

Clearly, the experimental results – surveying only upon undergraduate science students – do not distinguish between different groups of assessment audiences. They are, however, consistent with the existing literature on the use of probabilistic language, and they do suggest an important feature of these probability descriptors: that people both use and interpret them as containing information about event magnitude as well. People are more likely to choose more certain sounding probability descriptors (e.g., *likely* instead of *unlikely*) to discuss more serious consequence events. But people are also sensitive to this practice in others, expecting a certain amount of exaggeration about the likelihood of high magnitude events. A weather forecaster might describe a 10% probable snow flurry as *very unlikely*, which the television viewer would accurately interpret to mean about 10%. Likewise, a weather forecaster might describe a 10% probable hurricane as *medium likelihood*, which the television viewer would again accurately interpret to mean about 10%. The symmetry of the two groups allows for effective communication. Figure 4a illustrates this pattern. Assigning a fixed probability scale to describe uncertain events with significantly different magnitudes of impact could disrupt that symmetry, as seen in Figure 4b. What would happen if forecasters were to use a single phrase, such as *unlikely*, to describe both the hurricane and snowfall? Attempting to correct for the assumed exaggeration, the viewers would understand the single word *unlikely* as implying a smaller chance for the hurricane than for the snow flurries.

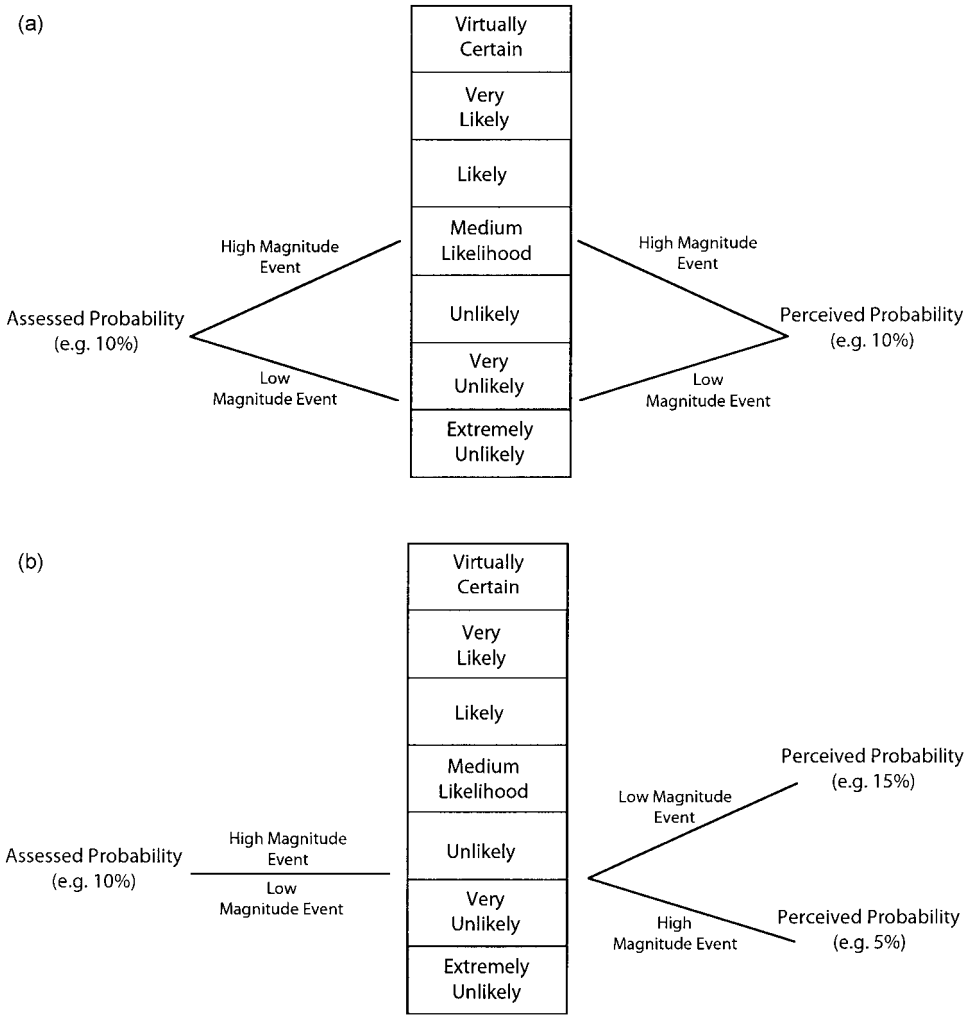

Figure 4. (a) Exaggeration and decoding. (b) Fixed scale and decoding.

5.1. BIASED MITIGATION EFFORTS

In response to the fixed probability scale, people will have a tendency to over-estimate the likelihood of low-magnitude events, and under-estimate the likelihood of high-magnitude events. Importantly, the two errors do not balance each other out, but introduce a bias in people's aggregate responses to the two events. Imagine, for example, that the hurricane, if it hits Boston, will cause damages of $10 million. The probability of this outcome is 10%, yielding an expected loss of $1 million, but people underestimate this probability to be 5%, yielding an expected loss of $0.5 million. The snow-flurries will cause very small damages, perhaps one additional road accidents costing $10,000. The probability is 10%, yielding an

expected loss of $1000, but people overestimate the probability to be 15%, yielding an expected loss of $1500. The underestimate of damages for the high-magnitude event completely overshadows the overestimate from the low-magnitude event. People's expectation of damages from the two combined events will be biased downward.

The efficiency of people's efforts to reduce damages, through advance preparation, will also be biased downward, with a net loss in welfare. To see how this is so, consider one possible strategy an individual or local area might pursue: the purchasing of insurance. First, imagine that it is possible to insure against each event at an actuarially fair rate, i.e., 10% of the possible loss from each event. Rational risk-averse actors would gain the greatest expected benefit from fully insuring against each event, purchasing $1 million of coverage for the hurricane, and $10,000 of coverage for the snow flurries, reducing to zero the variance of possible outcomes while leaving the expected outcome unchanged. But if people believed the probability of the hurricane were 5%, the insurance at a 10% rate would appear overpriced, and they would underinsure, i.e., purchasing insurance to cover $< \$1$ million. Likewise, estimating the likelihood of snow-flurries at 15%, people would over-insure. In each case, they would have purchased the wrong amount of insurance, resulting in positive variance, and a lowering of expected utility, for each event. Second, imagine that it is possible to purchase a single insurance policy for cover both events. At an actuarially fair rate of 10%, this policy would cost slightly more than $1 million. With the two errors in probability understanding, people would estimate losses at slightly more than $0.5 million. The policy would appear too expensive, and people would purchase less than full coverage.

5.2. THE IPCC STRATEGY

Climate change will bring many predictable impacts such as a rise in mean annual temperature, changing precipitation patterns, or mild coastal flooding. It also may bring less probable, more extreme impacts such as major coastal flooding (if polar ice were to deteriorate quickly), prolonged regional droughts, or large increases in storm frequency or intensity. Ideally, policies to mitigate and adapt to climate change will rely on an unbiased appraisal of both the probability and magnitude of each of these different possible outcomes. The communication strategy that the IPCC Third Assessment Report adopts – referring to probabilities through descriptive language matched to precise probability ranges – at first seems to be the best possible approach. Not only does it allow the IPCC more easily to achieve consensus within their own ranks about how to describe levels of confidence, but it also provides a lay audience with information that they can more easily digest.

At closer inspection, however, the strategy could be introducing an unintended bias into the policy process, namely one of under-responding to the aggregate risks associated with climate change. A careful reading of the report, in which the reader takes pains to note the precise probability ranges for each potential outcome, would

avoid such a bias. Many readers, however, may lack the time to read the report so carefully. Bias could enter in when readers make intuitive judgements about the likelihood of events, based on less attentive reading in which they fail continuously to match words with probability ranges.

Assessors can take steps to address this bias. If policy-makers read the report with attention to detail, they will both notice and adopt the IPCC's precise, potentially counterintuitive, meaning of probabilistic language. Scientists and assessors hence need to encourage the practice of careful reading, in particular highlighting the meaning of the probabilistic language, and not counting on the audience to do so on their own. But there are also steps that scientists can take to make sure that this happens. Most importantly, scientists should be aware that the potential for bias exists when an audience makes intuitive judgement. When communicating with policy makers or the lay public, scientists should encourage attention to detail. Whenever possible, scientists should refer to uncertainty with greater specificity than the report provides. Scientists should use not only the descriptive language of the report, but also matching those words to their respective probability ranges. As Moss and Schneider (2000) suggest, one approach could be to incorporate the uncertainty into decision-analytic frameworks, such as that carried out above for the simplified choice about purchasing insurance. Putting the numbers to use in this way encourages quantitative rigor, and through this rigor the audience can better understand the relative importance of the different potential outcomes of climate change. From a normative standpoint, the risks associated with low-probability high-magnitude events may be the most important elements of a rational decision-making framework addressing climate change. However, unless scientists encourage quantitative rigor on the part of policy-makers, it is likely the policy-makers will not give enough attention to these risks, and will take inadequate steps either to avoid or to prepare for these risks.

6. Conclusion

The strategy of using specifically defined language to describe the probabilities of climate change risks achieves important objectives, but may also introduce bias into policy-makers responses. Intuitively, people use such language to describe both the probability and magnitude of risks, and they expect communicators to do the same. Assessors need to emphasize that the IPPC's use of this language departs from people's expectations. Unless policy-makers appreciate this fact, their response to the assessment is likely to be biased downward, leading to insufficient efforts to mitigate and adapt to climate change.

References

Breyer, S.: 1993, *Breaking the Vicious Circle: Toward Effective Risk Regulation*, Harvard University Press, Cambridge.

Brun, W. and Teigen, K. H.: 1988, 'Verbal Probabilities: Ambiguous, Context-Dependent, or Both?', *Organizational Behavior and Human Decision Processes* **41**, 390.

Covello, V.: 1990, 'Risk Comparisons and Risk Communication: Issues and Problems in Comparing Health and Environmental Risks', in Kasperson and Stallen (eds.), *Communicating Risks to the Public: International Perspectives*, Kluwer Academic Publishers, Dordrecht, pp. 79–124.

Fischhoff, B.: 1996, 'Public Values in Risk Research', *Annals of the American Academy of Political and Social Science* **545**, 75.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., and Xiaosu, D. (eds.): 2002, *Climate Change 2001: The Scientific Basis*, Cambridge University Press, Cambridge U.K.

Irwin, A. and Wynne, B. (eds.): 1996, *Misunderstanding Science?*, Cambridge University Press, Cambridge U.K.

Kahneman, D. and Tversky, A.: 1979, 'Prospect Theory: An Analysis of Decision under Risk', *Econometrica* **47**, 263.

Kammen, D., Shlyakhter, A., and Wilson, R.: 1994, 'What is the Risk of the Impossible?', *Journal of the Franklin Institute* **331A**, 97.

Leiss, W.: 1996, 'Three Phases in the Evolution of Risk Communication Practice', *Annals of the American Academy of Political and Social Science* **545**, 85.

Morgan, M. G. and Keith, D. W.: 1995, 'Subjective Judgments by Climate Experts', *Environmental Science and Technology* **29**, 468A.

Moss, R. H. and Schneider, S. H.: 2000, 'Uncertainties in the IPCC TAR: Recommendations to Lead Authors for more Consistent Assessment and Reporting', in Pachauri, R., Taniguchi, T., and Tanaka, K. (eds.), *IPCC Supporting Material, Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, pp. 33–51.

Patt, A.: 1999, 'Assessing Extreme Outcomes: The Strategic Treatment of Low Probability Impacts in Scientific Assessment', *Risk Decision and Policy* **4**, 1.

Patt, A. and Gwata, C.: 2002, 'Effective Seasonal Climate Forecast Applications: Examining Constraints for Subsistence, Farmers in Zimbabwe', *Global Environ. Change* **12**, 185.

Patt, A. and Zeckhauser, R.: 2002, 'Behavioral Perceptions and Policies toward the Environment', in Gowda, R. and Fox, J. (eds.), *Judgments, Decisions, and Public Policy*, Cambridge University Press, Cambridge U.K., pp. 265–302.

Pinker, S.: 1997, *How the Mind Works*, Norton, New York.

Van der Sluijs, J.: 1997, *Anchoring Amid Uncertainty, on the Management of Uncertainties in Risk Assessment of Anthropogenic Climate Change*, Ludy Feyen, Leiden, The Netherlands.

Wallsten, T. S., Fillenbaum, S., and Cox, J. A.: 1986, 'Base Rate Effects on the Interpretation of Probability and Frequency Expressions', *Journal of Memory and Language* **25**, 571.

Weber, E. U.: 1994, 'From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events', *Psychological Bulletin* **115**, 228.

Weber, E. U. and Hilton, D. J.: 1990, 'Contextual Effects in the Interpretations of Probability Words: Perceived Bas Rate and Severity of Events', *Journal of Experimental Psychology: Human Perception and Performance* **16**, 781.

Windschitl, P. D. and Weber, E. U.: 1999, 'The Interpretation of "Likely" Depends on the Context, but "70%" is 70% – right? The Influence of Associative Processes on Perceived Certainty', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **25**, 1514.

(Received 6 July 2001; in revised form 10 February 2003)

Climatic Change (2007) 85:19–31
DOI 10.1007/s10584-007-9315-7

Expressions of likelihood and confidence in the IPCC uncertainty assessment process

James S. Risbey · Milind Kandlikar

Received: 22 February 2005 / Accepted: 13 July 2007 / Published online: 6 September 2007

Abstract Communication of uncertainty information in Intergovernmental Panel on Climate Change (IPCC) assessments has evolved through successive reports to provide increasingly formal classifications for subjective and objective information. The first IPCC assessments provided uncertainty information in largely subjective form via linguistic categorizations depicting different levels of confidence. Recent assessments have codified linguistic terms to avoid ambiguity and introduced probabilistic ranges to express likelihoods of events occurring. The adoption of formal schemes to express likelihood and confidence provides more powerful means for analysts to express uncertainty. However, the combination of these two metrics to assess information may engender confusion when low confidence levels are matched with very high/low likelihoods that have implicit high confidence. Part of the difficulty is that the degree to which different quantities in the assessments are known varies tremendously. One solution is to provide likelihood information in a scheme with a range of different precision levels that can be matched to the level of understanding. A version of this scheme is also part of the IPCC uncertainty guidance and is described here.

1 Introduction

Every assessment of climate change is faced with the need to characterize and communicate uncertainties in the state of understanding. This has long been a contentious process. For example, the 'Charney' report (National Research Council

J. S. Risbey (✉)
CSIRO Marine and Atmospheric Research, Hobart, Tasmania 7001, Australia
e-mail: james.risbey@csiro.au

M. Kandlikar
Institute for Asian Research, University of British Columbia, Vancouver,
British Columbia V6T1Z2, Canada

1979) characterized climate sensitivity as 3°C ± 1.5°C, sparking a long running discussion of the meaning of the error range. As climate assessments moved into the Intergovernmental Panel on Climate Change (IPCC) process and attracted larger audiences, pressure has mounted to formalize the characterization of uncertainties. The IPCC has provided a forum to develop standard, useful formats for communicating uncertainty. The first IPCC reports highlighted subjective judgements by categorizing results according to various linguistic expressions of confidence: "we are certain of …", "we calculate with confidence that …", "our judgement is that". Yet, considerable ambiguity remained in the interpretation of these terms and in the lack of any formal method to present quantitative information and likelihoods.

Formal consideration of uncertainty in assessments of the IPCC began with the Third Assessment Report (TAR) (Houghton et al. 2001). As a part of the TAR process, Moss and Schneider (2000) wrote a guidance document that developed a framework for representing and communicating uncertainty in quantitative and qualitative terms. A critical aspect of the Moss and Schneider proposal was a scheme to link qualitative descriptors of uncertainty to quantitative metrics. They proposed a subjective five-unit scale that mapped quantitative ranges of subjective confidence to linguistic descriptors of confidence. They also recognized that in some cases, scientists might want to either supplement or supplant subjective quantitative judgements with descriptions of uncertainty that are qualitative in nature. Accordingly, they also developed the qualitative schema shown in Table 1, which differentiated between the quality of evidence and the level of expert consensus on a particular topic. The Moss and Schneider scale is Bayesian in that it does not make explicit the distinction between subjective and frequentist forms of uncertainty.

This distinction has emerged in representation of uncertainty in the IPCC fourth assessment (AR4) process. Guidance documents on uncertainty communication produced for the AR4 (IPCC 2006) distinguish between *likelihood* and *level of confidence*[1] in representations of uncertainty. In these documents 'likelihood' has a frequentist connotation and refers to a probabilistic assessment of some well-defined outcome having occurred or expected to occur in the future (see Table 2). Subjective expression of uncertainty is introduced in the AR4 via a 'level of confidence' scale. The 'level of confidence' is based on the degree of understanding in the expert community using a probabilistic formulation (see Table 3). In traditional use, likelihoods are based on existing data (from observations or models), and can be applied to all cases where the classical definition of probability based on past counts applies. Levels of confidence can be applied when such data are incomplete and subjective judgement is required.

Separating the likelihood of, and the level of confidence in, a statement has also raised the possibility that likelihood and level of confidence could be combined in making assessments. Indeed, guidance documents for the AR4 assessment of uncertainty state that "it is possible to have high confidence in a finding indicating that climate change would lead to a low probability of some outcome and conversely to have low confidence in a finding that climate change would lead to a high probability of another outcome" (Manning and Petit 2004). While it is possible to combine likelihood and confidence estimates in this way, it is not necessarily

[1]The term 'level of confidence' is distinct from that of a 'confidence interval' as normally used in statistics.

Table 1 This table is reproduced from Moss and Schneider (2000) and provides a qualitative means to express levels of confidence based on the level of expert agreement and the relevant evidence

	Amount of evidence →	
Level of agreement ↑	Established but incomplete Speculative	Well established Competing explanations

The two-by-two matrix allows 'low' or 'high' determinations of consensus and evidence to determine the appropriate confidence expression.

meaningful in all cases. In the following section we explore a variety of possible ways to combine likelihoods and level of confidence. The combinations we consider are using only likelihood, using only confidence levels, and using the two together.

The focus on uncertainty in this paper is on characterizing uncertainty in 'outcomes'. By this we mean outcomes of some event or well-posed question that are subject to quantification or can be expressed in probabilistic form. What is important in this case is that the event and variable be well specified. We are not concerned here about characterizing the event according to uncertainty typologies, only in expressing the outcome. Similarly, we are not concerned here with the assessment of uncertainty in more general knowledge claims which do not have quantifiable metrics. We assess each of the likelihood and confidence schemes below only in terms of their utility for expressing quantifiable outcomes.

2 Likelihood versus levels of confidence: three alternatives

On the face of it, the distinction between likelihood and level of confidence sounds unproblematic and almost innocuous. We argue however, that care must be taken in applying these differing definitions if the goal of effective communication is to be achieved, especially when they are used in combination. In what follows, we examine pros and cons of using likelihood and levels of confidence individually and in combination by defining three alternative options for using likelihood and levels of confidence in communication of uncertainty. We then go on to suggest one alternative for how the two approaches can be fruitfully used in combination.

2.1 Alternative 1: use only likelihood

Alternative 1 would simply use likelihoods and can be interpreted in two ways. The first is the frequentist/classical view and ignores or downplays subjective uncertainty.

Table 2 Likelihood defined as a probabilistic assessment of some well-defined outcome having occurred or occurring in the future

Terminology	Likelihood of occurrence (%)
Virtually certain	> 99
Very likely	> 90
Likely	> 66
About as likely as not	33 to 66
Unlikely	< 33
Very unlikely	< 10
Exceptionally unlikely	<1

Table based on IPCC (2006).

Table 3 Characterizations of levels of confidence expressed in terms of the odds of being correct

Terminology	'Odds' of being correct
Very high confidence	At least 9 out of 10 chance
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very low confidence	Less than 1 out of 10 chance

Table based on IPCC (2006).

On the plus side, this is clear and works well for traditional science problems (Ravetz 1971), especially those based on empirical observations. It also provides scientists with a greater degree of comfort by making a 'clean' separation between objective and subjective knowledge [assuming this were possible (Schrader-Frechette 1984)]. Any information that does not confirm to the norms of classical statistics is discarded in this approach. Unfortunately, this alternative does not always work well for climate change because only a limited set of quantities can be expressed in full likelihood terms. This is effectively the pre-IPCC approach where uncertainty was discussed in as much as rigorously quantifiable measures were available (and avoided if not). Using only classical likelihoods delimits the boundaries of knowledge to a small set of findings based either on the historical record or from models rigorously calibrated to historical data. This makes it difficult to provide meaningful scientific advice to policy makers on a large number of questions where history is unlikely to be a suitable guide for the future.

The second view of likelihood assessments acknowledges that any likelihood assessment would contain subjective elements. With this broader view, one need not limit the domain of applicability of likelihood to problems rich in past count data. However, the more subjective the likelihood assessment, the more the need to evaluate that subjectivity, and the more the assessment would be improved by adding information on confidence levels as well. One way to reduce subjective and uncertain elements of a likelihood assessment is to render these elements conditional—declare assumptions and hold them fixed. For example, one may express likelihoods of change conditional on a particular emissions scenario. This can increase the utility of likelihood statements, though one is still left with a need to assess and communicate the quality of the conditionals. This last step is not very amenable to a likelihood-only scheme.

In summary, if we view likelihoods in strict frequentist terms, they have limited application in addressing climate change issues. If on the other hand we imbue likelihoods with subjective content (or express them conditionally), then they have wider application in climatology, but that subjectivity (conditionality) needs to be evaluated. However, the likelihood scheme itself is inappropriate for subjective evaluations and needs to be supplemented with a qualitative framework.

2.2 Alternative 2: use only levels of confidence

As noted above, it is rare to have rigorous likelihood data for all but a few variables. The likelihood of future events can be formally determined from models, but models have many subjective elements (van der Sluijs et al. 1998; Shackley et al. 1999; Murphy et al. 2004). Thus, it could be argued that a majority of data relevant to

assessment of future climate change has embedded subjective elements. The quality of the data, and suitability of models for the questions posed determines the level of confidence (Risbey 2002). The early IPCC reports (Houghton et al. 1990, 1996) took a level of confidence approach without quantification. Instead they used linguistic approaches, employing terms such as "we calculate with confidence that . . . ", without providing quantitative measures. The problems of linguistic assessment without quantification are well known (Wallsten et al. 1993). The TAR uncertainty guidance document (Moss and Schneider 2000) changed this by encouraging scientists to assign a quantitative scale to linguistic claims. The Moss and Schneider scheme was partially adopted in the TAR, and provided a forceful starting point for effective communication of uncertainty.

There are however reasons for the IPCC to extend uncertainty communication beyond the level of confidence scale. Reasonable argument can be made objecting to the use of a purely subjective scale. For one, uncertainty surrounding a few key variables in the historical record (e.g. global mean temperature change) can be assessed in a largely objective manner. In such cases, it probably makes practical sense to preserve the distinction between objective and subjective assessment. Second, some members of the scientific community might, for a variety of reasons, not be comfortable using an exclusively Bayesian approach (Giles 2002). Since the IPCC is ostensibly a consensual scientific organization it takes intellectual pluralism seriously, and accommodates differing but valid perspectives on uncertainty. Hence, using levels of confidence along with likelihood provides a useful way of combining different levels of knowledge, while satisfying the needs of a consensus process. We turn next to the question of how likelihood and confidence can best be used in combination.

2.3 Alternative 3: use levels of confidence and likelihood

In this section we discuss two schemes for combining measures of both likelihood and confidence. The first scheme uses both measures together to condition likelihoods by level of confidence. The second scheme adjusts the metric used to express likelihood according to the level of confidence.

2.3.1 Alternative 3a: simultaneous use of likelihood and levels of confidence

The simplest possible approach is to simultaneously combine likelihood and levels of confidence in communicating uncertainty. For instance, the AR4 uncertainty guidance document allows provision for likelihood to be used to communicate the probability/variability in a particular outcome (Table 2), and for the level of confidence to communicate the level of agreement associated with that likelihood (Table 3). The logic of this approach appears to be reasonable—provide likelihoods based on available data, but also communicate a subjective uncertainty about the likelihood. Upon closer inspection it becomes clear that simultaneous use of likelihood and level of confidence can cause confusion and make the already difficult challenge of communicating uncertainty even more difficult. Below we provide a description of why this may be the case.

All likelihood outcomes (of high, medium or low likelihood) with a low subjective confidence cannot be interpreted in a quantitative manner. Part of the problem with

the approach is that likelihood and confidence cannot be fully separated. Likelihoods contain implicit confidence levels. When an event is said to be extremely likely (or extremely unlikely) it is implicit that we have high confidence. It wouldn't make any sense to declare that an event was extremely likely and then turn around and say that we had low confidence in that statement. For example, if we declare that it is extremely likely to rain tomorrow, but then say that we have very low confidence in that statement, that would lead to a state of confusion. People would rightly ask us how we could give such a high (near certain) likelihood to an event about which we profess to have little understanding. If we say there is a 99% chance of rain, that implies that we are nearly certain it is going to rain, which means that we must have high confidence, never low.

As we show in Table 4, interpreting uncertainty when there are two levels of imprecision is in some cases rather difficult. The table shows likelihoods conditioned by levels of confidence. First consider a statement whose likelihood is high, and subjective confidence on this likelihood is high. This entry in the table is easy to interpret and would have high likelihood. In fact, all entries in the row related to high confidence are easy to interpret. The likelihood gives an estimate of the probability of some event occurring, and the high confidence estimate tells us that the subjective error bars for that estimate are small. Thus, the likelihoods conditioned by confidence level are equivalent to the likelihoods when the level of confidence is high.

Now consider entries in the row related to low confidence, and we run into difficulties of interpretation. If low confidence translates into large error bars about the likelihood estimate, then the actual likelihood (could it be known) may bear little relationship to the estimated likelihood. Further, we encounter real problems when combining very high or very low likelihood estimates (the 'virtually certain' and 'exceptionally unlikely' from Table 2) with low confidence assessments. As pointed out earlier, very high or very low likelihoods are only meaningful when confidence is high. By allowing low confidence assignments to such estimates, confusion may be created. The simultaneous likelihood/confidence scheme allows the analyst to create contradictory combinations of likelihood and confidence.

Combinations of likelihood and confidence with medium levels of confidence are intermediate between the high and low confidence cases, which makes them somewhat ambiguous. If medium confidence implies moderate error bars, then the

Table 4 Likelihoods conditioned by levels of confidence

		likelihood		
		low	medium	high
confidence	low	?	?	?
	medium	low–med	low–high	med–high
	high	low	medium	high

This table shows the simultaneous use of likelihoods and levels of confidence. Confidence and likelihood levels are each classified into 'low', 'medium' and 'high'. Entry in each element of the table represents how the associated confidence level modifies the likelihood. For example, when the likelihood of an outcome is low and subjective confidence levels about the science surrounding that outcome are high, then the likelihood is low. However, when likelihood of an outcome is high and the subjective confidence levels associated with that outcome are low it is impossible to interpret the likelihood in a meaningful way. Such combinations that are difficult to interpret are represented by a '?'.

likelihood estimates may be reasonable, but perhaps one category too high or low. In Table 2 then, the low likelihood estimate might correspond to a confidence conditioned likelihood in the range from low to medium likelihood. And the high likelihood estimates could correspond to conditioned likelihoods from medium to high for example. That is, medium confidence implies a spreading of the likelihood ranges, assuming that the error bars are moderate. However, if medium confidence implies larger error bars, then the same confusion that applies to the low confidence cases would apply to some degree to the medium confidence cases.

A scheme that combines estimates of confidence and likelihood is increasingly difficult to interpret the lower the estimate of confidence. In the extremes at low confidence and very high or low likelihood, the combinations make little sense. These features of the scheme would create a conundrum for analysts that may lead them to avoid low and medium confidence combinations. This could result in a bias toward expressing results with higher confidence, since it is meaningful with this scheme to present only statements associated with higher confidence. Thus an uncertainty scheme that simultaneously uses likelihood and confidence is ripe to either contradict itself or bias towards suppression of low confidence. Hence, we argue that simultaneous use of likelihood and levels of confidence can be dangerous. Below we propose a method that merges some of the positive aspects of both approaches (alternatives 1 and 2).

2.3.2 Alternative 3b: adjust likelihood scale according to warranted precision

Another approach to using both likelihoods and levels of confidence is to use a progressive scheme that articulates the basis for the assessment of each attribute (Risbey et al. 2002; Kandlikar et al. 2005). This scheme allows analysts to use a sequential process that does not treat all uncertain variables as statistically quantifiable, and provides a mechanism for communicating uncertainty at a level appropriate to existing scientific understanding. The sequential process is outlined in Table 5 and described below. The process begins by asking the analyst if a probability distribution for the outcome or variable under consideration can be provided (i.e., full likelihood information). This serves to capture either those variables for which historical data exists, or those for which there is sufficient consensus. If a pdf can be given then one moves on to the next variable of interest. For many/most variables this is not the case however, so it is necessary to have more coarse means of representing the uncertainty as well. In these cases the analyst moves down to the next level in the

Table 5 Characterizations of likelihood for a graduated range of precision levels ranging from fully specified probability distributions through to qualitative declarations of knowledge and ignorance

	Measure of likelihood	Justification
1	Full probability density function	Robust, well defended distribution
2	Bounds	Well defended percentile bounds
3	First order estimates	Order of magnitude assessment
4	Expected sign or trend	Well defended trend expectation
5	Ambiguous sign or trend	Equally plausible contrary trend expectations
6	Effective ignorance	Lacking or weakly plausible expectations

scheme. At each level down the degree of quantification (precision) is reduced. The idea is to express quantities at a level of precision commensurate with the degree of confidence with which the quantity is known. The steps in the scheme are as follows:

Step 0: Definition Define the variable or outcome to be examined and the context in which it is being examined. Though seemingly trivial, this first step is crucial in ensuring that the outcome in question has a commonly shared understanding and can be meaningfully quantified. This step also facilitates comparison of uncertainties across studies and through time.

Step 1: Full probability density function (Robust, well defended probability distribution): Is it reasonable to specify a full probability distribution for the outcome? If yes, specify the distribution. Justify your choice of distribution and 5th and 95th percentiles. Are there any processes or assumptions that would cause the 5/95 percentiles to be much wider than you have stated? This is the full likelihood description. If you cannot provide justifications for why you consider the distribution shape and 5th and 95th percentiles to be fairly robust, then move to a lower precision category (step 2).

Step 2: Bounds (Well defended bounds): Is it reasonable to specify bounds for the distribution of the outcome? If yes, specify 5th and 95th percentiles. Can you describe any processes or assumptions that could lead to broader/narrower bounds? If so, describe and revise. The choice of 5th/95th percentiles is by convention. Other ranges (e.g. 10th/90th) could also be used by different research communities as long as the choice is made clear. If the bounds are robust to assumptions, then specify your 5/95 bounds and your reasoning for placing them where you did. If you cannot provide bounds confidently then go to step 3.

Step 3: First order estimates (Order of magnitude assessment): If appropriate, specify and justify your choice of a first order estimate for the value of the variable, indicating the main assumptions behind the value given. In specifying a value, do not report more precision than is justified. For example, if the value is only known to a factor of two or an order of magnitude, then report it in those terms. In some cases, powers of ten may be appropriate; in other cases more nuanced scales may be used so long as they are declared and supported. How robust is your estimate to underlying assumptions? If it is not particularly robust to the set of assumptions or outcomes you listed, then go to step 4.

Step 4. Expected sign or trend (Well defended trend expectation): While it may not be possible to place reliable bounds or a magnitude on the expected change in a variable, you may still know something about the likely trend. Can you provide a reasonable estimate of the sign or trend (increase, decrease, no change) of the expected change? If so, give the expected trend and explain the reasoning underlying that expectation and why changes of the opposite sign or trend would generally not be expected. Describe also any conditions that could lead to a change in trend contrary to expectations. It is reasonable to include in this category changes which have a fair degree of expectation, but which are not certain. The distinction between this category and the following one is that the arguments for the expected change

should be significantly more compelling or likely than those for a contrary change. If the arguments tend towards a more equal footing, then step 5 (ambiguous sign) is more appropriate.

Step 5: Ambiguous sign or trend (Equally plausible contrary trend expectations): In many cases it will not be possible to outline a definitive trend expectation. There may be plausible arguments for a change of sign or trend in either direction. If that is the case, state the opposing trends and outline the arguments on both sides. Note key uncertainties and assumptions in your arguments and how they may tip the balance in favour of one trend direction or the other. If information about the variable does not support this kind of supposition, then go to step 6.

Step 6: Effective ignorance (Lacking or weakly plausible expectations): In most cases we know quite a bit about the outcome variable. Yet despite this, we may not know much about the factors that would govern a change in the variable of the type under consideration. As such, it may be difficult to outline plausible arguments for how the variable would respond. If the arguments used to support the change in the variable are so weak as to stretch plausibility, then this category is appropriate. Selecting this category does not mean that we know nothing about the variable. Rather, it means that our knowledge of the factors governing changes in the variable in the context of interest is so weak that we are effectively ignorant in this particular regard. If this category is selected, describe any expectations, such as they are, and note problems with them.

These six steps provide a mechanism for making explicit the reasons for low/high levels of confidence based on assessments of data quality and scientific knowledge at each step. Responses can be given in progressively relaxed quantitative forms, ranging from full likelihood form (when justified) through to more qualitative characterizations as appropriate. The analyst moves down through the steps and stops when the level of confidence in the variable matches the precision available in the category. Though the method is subjective, it is transparent in that it asks the analyst to provide justifications for the form of quantification selected. Thus the reasoning is clear and explicit for others to scrutinize. The approach provides a simple, yet consistent way to use likelihood information in conjunction with subjective knowledge.

In this scheme, the form (scale) in which likelihoods are expressed is conditioned by subjective judgement (confidence) such that likelihood and confidence remain consistent with one another at low confidence levels—likelihoods are expressed in coarse quantitative form when confidence is low. This contrasts with the simultaneous likelihood/confidence scheme where likelihood is conditioned by level of confidence, but the form of likelihood expression does not change.

The determination of appropriate precision levels to express outcomes proceeds through a process of argumentation that progressively excludes over-precise and under-precise levels. In practice that determination won't always be obvious, and the analyst may wish to employ a variety of measures to assess the quality and precision of the outcome variable. One approach which appears to be well suited is to use the NUSAP scheme (Funtowicz and Ravetz 1990), which employs methods to determine the 'pedigree' and quality of the relevant data and methods used (e.g. van der Sluijs

et al. 2005a,b). Any method which helps to determine the appropriate precision of outputs could be used in this regard.

3 Likelihood example

A simple example may be helpful in illustrating use of the likelihood schemes. Suppose we are asked what the likelihood is of the thermohaline circulation shutting down in response to increased greenhouse forcing of the climate (Mastrandrea and Schneider 2004). First, we need to remove any ambiguity in the question, so we would need to specify a time range over which this event might occur. In the limit to infinity, the question would be almost trivial as the circulation has shutdown in the past and would likely do so at some point in the future, greenhouse warming or not. Thus, we might limit the period to some point in time such as 2100 or until some particular CO_2 concentration is reached, say 550 ppm. Further, we need to define whether the flow simply slows down or reverses completely. Suppose we rephrase the question then as to what is the likelihood that the thermohaline circulation reduces by at least half for a stabilized CO_2 concentration of 550 ppm? Of course one would also need to specify or factor in the time path of greenhouse and other emissions and the climate sensitivity to make the question more precise and account for further uncertainties. Since we only consider hypothetical responses to the question by way of example, the full specification of uncertainties is not critical here.

In the case that the likelihood was very high that the circulation would be reduced by at least half, the answer could be phrased easily enough with either scheme. In the combined likelihood/confidence scheme (scheme 3a) one would use the likelihood scale of Table 2 to yield the answer 'virtually certain'. In order to make such a certain determination, confidence would also have to be very high. In the precision-based scheme (scheme 3b), if the underlying knowledge about thermohaline circulation responses to CO_2 were very robust, one could specify a probability distribution for the value of the flow at 550 ppm. From the distribution it would be clear in this case that the vast bulk of the probability mass favoured flows less than half the present value.

Suppose now that confidence in the assessment of thermohaline circulation changes was not high, but low. Following the earlier quote of Manning and Petit (2004), could one meaningfully specify that the outcome was virtually certain with low confidence? We suspect that this would be interpreted by many to mean that the likelihood of the likelihood (virtually certain) was low; i.e. that it was not likely to be in this category. But that is presumably not exactly what is intended, for one would do much better in that case to simply specify a more likely likelihood category. Further, the specification of likelihoods upon likelihoods conjures up the notion of an infinite regress (Funtowicz and Ravetz 1990). Presumably, what is intended here is that one has low confidence in specifying the likelihood and really means to say that the likelihood is unknown. There is no point in specifying a very precise category ($>$ 99% chance) and then saying, "well, we don't know much about that". This would be true of every likelihood category in this case and it could be misleading to single out a single category and make the statement. Rather, one wishes to convey the uncertainty about likelihood directly and without ambiguity.

In the precision-based scheme, if confidence is low there is no point trying to provide a pdf or even percentile bounds on the value of the flow at 550 ppm. One

moves down through the scheme to find an expression for the expected changes in the flow that matches the level of understanding about that. If a 'first-order estimate' could reasonably be given for the value of the flow at 550 ppm, that would be specified, together with declarations of assumptions underlying the value given. If confidence were too low to warrant that, one might simply specify the 'expected sign' for changes in the flow. There are good reasons to expect a decrease in the flow (Manabe and Stouffer 1993), and this much at least could be conveyed, along with the reasoning. This is not a likelihood in that it does not answer the question in terms of a probability of the specified outcome. But since that probability would be vacuous, it doesn't make sense to give one. One provides as much quantifiable information as the level of confidence will support. If confidence were even lower again, the alternative scheme (3b) allows for more speculative declarations of knowledge and ignorance. In the fixed-precision likelihood scheme of Table 2, one cannot retreat from specifying a probability, no matter how tenuous the knowledge base. The fixed-precision likelihood scheme thus becomes a straight-jacket for the analyst when uncertainties increase.

4 Conclusions

The sequence of quinquennial IPCC reports from 1990 to the present time provides an interesting study of the evolution of formal uncertainty communication in the climatological community. In the period prior to the first IPCC reports, subjective information about uncertainty tended to be included in ad hoc ways. Some of the first reports on climate change, such as National Research Council (1979) provided discussion of the reasonings that led them to select particular ranges (for climate sensitivity for example), though ambiguity remained about just what the ranges were supposed to represent (van der Sluijs 1997). Subsequent climate change assessments used a range of different styles to communicate qualitative and quantitative dimensions of uncertainty (National Research Council 1982, 1992; Jaeger 1988).

The earliest IPCC reports provided qualitative statements of confidence in expressing results. This process was formalized by Moss and Schneider (2000) for the third IPCC assessment. They introduced a formal scale for assessing levels of confidence, which was adapted by IPCC (2006) to express confidence in terms of the odds of being correct (Table 3). With preparation for the fourth IPCC assessment, formal expressions of confidence have now been supplemented with formal expressions of likelihood (IPCC 2006) (Table 2). This provides analysts with the ability to describe both the chance of some outcome occurring and the confidence they have in their prediction of that chance. This offers greater flexibility over uncertainty expression schemes based on only one of these dimensions. With *confidence only* schemes there is too little attention to the extraction of appropriate likelihood information for risk assessments (Dessai and Hulme 2004) and inevitable ambiguity when likelihood information is given. Schemes based on *likelihood only* fall short because they assume that all relevant uncertainty information can be communicated in likelihood terms. In practice there are large subjectivities underlying many climate assessments that are best addressed through some form of confidence assessment.

Recognizing the advantages of using both likelihood and confidence information, the IPCC AR4 has provided schemes for both these concepts in communicating uncertainties (Allen et al. 2004; IPCC 2006; Manning 2006). Some analysts will

undoubtedly use these two schemes in combination. However, some combinations of likelihood and confidence (as these concepts are defined by the IPCC) are difficult to interpret. The source of the problem is that likelihood levels contain implicit confidence levels. For example, very high/low likelihoods only seem meaningful if interpreted as statements of high confidence. Yet a simple combination of the IPCC AR4 likelihood and confidence schemes theoretically allows analysts to create confusing combinations of low confidence and high likelihood. Such combinations will usually be avoided. However, by avoiding the confusing combinations in this scheme the analyst may bias uncertainty communication by under-reporting low confidence cases. This also has the effect of reducing the likelihood/confidence scheme back to a *likelihood only* scheme since only higher confidence cases are retained.

A version of the alternative scheme for reporting uncertainty given here has also been incorporated into the AR4 uncertainty guidance. The scheme described here starts with the recognition that different quantities dealt with in IPCC reports are known with differing levels of precision. One-size-fits-all precision schemes are bound to be left wanting in such circumstances. The expression of likelihood needs to be flexible enough to take into account a range of precision levels from fully quantitative pdf's to virtual ignorance of quantitative changes. Rather than having a single quantitative format with a single precision for expressing likelihood, the level of precision is relaxed as the underpinning knowledge of the quantity degrades. Full quantitative information is provided for likelihoods when it is reasonable to do so, and only then. Confidence enters into the alternative scheme via choice of precision level and by way of explanation; not formally through a labelling scheme like Table 3. That is, one has to defend the choice of level of precision, explaining reasoning, outlining assumptions, and evaluating the robustness of the choice. The levels of confidence are implicit in the choices of quantitative category and in the articulation of the factors underlying those choices. The assessment of confidence thus conditions the form in which likelihood is expressed, rather than the value itself as in conventional schemes.

The IPCC's uncertainty assessments are marked by increasing formalism, thus reducing linguistic sources of ambiguity. These developments in uncertainty communication codified in the IPCC provide a richer platform to communicate climate science for policy, though potential for confusion remains. The new formalisms are beginning to incorporate deeper forms of uncertainty, opening the door for more pluralistic conceptions of uncertainty in future assessments.

Acknowledgements This work was funded by the Australian Research Council and the Wealth from Oceans Flagship program. Helpful comments were provided by Suraje Dessai, Chris Forest, and Jeroen van der Sluijs.

References

Allen M, Booth B, Frame D, Gregory J, Kettleborough J, Smith L, Stainforth D, Stott P (2004) Observational constraints on future climate: distinguishing robust from model-dependent statements of uncertainty in climate forecasting. In: IPCC workshop on describing scientific uncertainties in climate change to support analysis of risk and of options, Ireland, 11–13 May 2004

Dessai S, Hulme M (2004) Does climate adaptation policy need probabilities? Clim Policy 4(2):107–128

Funtowicz S, Ravetz J (1990) Uncertainty and quality in science for policy. Kluwer, Dordrecht, p 229
Giles J (2002) When doubt is a sure thing. Nature 418:476–478
Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds) (2001) Climate change 2001: the scientific basis. Cambridge University Press, Cambridge, p 881
Houghton JT et al (eds) (1996) Climate change 1995: the science of climate change. Cambridge University Press, Cambridge, p 572
Houghton JT, Jenkins G, Ephraums J (eds) (1990) Climate change: the IPCC scientific assessment. Cambridge University Press, Cambridge, p 365
IPCC (2006) Guidance notes for lead authors of the IPCC fourth assessment report on addressing uncertainties. Appendix in Manning 2006. Adv Clim Change Res 2:13–21
Jaeger J (1988) Developing policies for responding to climate change. Technical report, World Meteorological Organization. WCIP-1, WMO/TD-No. 225, p 53
Kandlikar M, Risbey J, Dessai S (2005) Representing and communicating deep uncertainty in climate change assessments. Comptes Rendus Geosci 337(4):443–455
Manabe S, Stouffer R (1993) Century-scale effects of increased atmospheric CO_2 on the ocean-atmosphere system. Nature 364(6434):215–218
Manning M (2006) The treatment of uncertainties in the Fourth IPCC Assessment Report. Adv Clim Change Res 2(1):13–21
Manning M, Petit M (2004) A concept paper for the AR4 cross cutting theme: uncertainties and risk. Position paper for the IPCC Risk and Uncertainty workshop, Maynooth, Ireland, May 2004
Mastrandrea M, Schneider SH (2004) Probabilistic integrated assessment of 'dangerous' climate change. Science 304:571–575
Moss R, Schneider SH (2000) Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. In: Pachauri R, Taniguchi T, Tanaka K (eds) Guidance papers on the cross cutting issues of the third assessment report of the IPCC. Technical report. World Meteorological Organization, Geneva, pp 33–51
Murphy J, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772
National Research Council (1979) Carbon dioxide and climate: a scientific assessment. National Academy Press, Washington, DC, p 22
National Research Council (1982) Carbon dioxide and climate: A second assessment. National Academy Press, Washington, DC, p 72
National Research Council (1992) Policy implications of greenhouse warming: mitigation, adaptation, and the science base. National Academy Press, Washington, DC, p 944
Ravetz J (1971) Scientific knowledge and its social problems. Clarendon, Oxford, p 449 (reprint: Transaction, New Brunswick, 1996)
Risbey JS (2002) Comment on Soon et al. 2002: modeling climatic effects of anthropogenic carbon dioxide emissions: unknowns and uncertainties. Clim Res 22(2):185–186
Risbey JS, Lamb PJ, Miller RL, Morgan MC, Roe GH (2002) Exploring the structure of regional climate scenarios by combining synoptic and dynamic guidance and GCM output. J Clim 15(9):1036–1050
Schrader-Frechette K (1984) Science policy, ethics, and economic methodology: some problems of technology assessment and environmental impact analysis. Reidel, Dordrecht
Shackley S, Risbey J, Stone P, Wynne B (1999) Adjusting to policy expectations in climate change modeling: an interdisciplinary study of flux adjustments in coupled atmosphere-ocean general circulation models. Clim Change 43(2):413–454
van der Sluijs J (1997) Anchoring amid uncertainty. Ph.D. Thesis, University of Utrecht, Utrecht, p 260
van der Sluijs J, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J (2005a) Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. Risk Anal 25(2):481–492
van der Sluijs J, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J (2005b) Experiences with the NUSAP system for multidimensional uncertainty assessment. Water Sci Technol 52(6):133–144
van der Sluijs J, Eijndhoven J, Shackley S, Wynne B (1998) Anchoring devices in science for policy: the case of consensus around climate sensitivity. Soc Stud Sci 28(2):291–323
Wallsten T, Budescu D, Zwick R (1993) Comparing the calibration and coherence of numerical and verbal probability judgments. Manage Sci 39:176–190