The development of 'for experts systems' as heuristic reasoning platforms in risk decision support: A consideration of tool design, technology transfer and compatibility with Bayesian decision analysis.

J.G. Arthur

December 2007

Acknowledgements

This thesis is dedicated at the emotional, intellectual and practical levels.

Emotionally:
To Helen Arthur, without whom more than just this would not have been possible.

Intellectually:
To Jim Smith, for being one of those rare people who makes me feel clever.
And to Henry Wynn for his early confidence.

Practically:
To George Gordon, Julia Sonander and Matryn Richards, for being the kind of people who can channel what I do into something worthwhile.

Abstract

This work considers the creation of two risk and decision support systems, one for the National Air Traffic Services of the UK and one for Unilever, a multi-national. Their development contributes to risk decision science in the area of decision support in particular. This contribution is based on the development real-life systems, it has three key elements. One, it addresses the fact that, for practical environments like these, the science of risk and decisions is insufficiently resolved to be accepted and easily used. Two, the systems share an arena with subjective Bayesian decision analysis. The benefits of a hybrid form of the two approaches to generate higher levels of user acceptance and organisational transfer is discussed. Three, they take the unique approach of being 'for experts' systems rather than 'expert systems'. This approach offers a number of benefits to applied user communities. These include: a decision support system which remains grounded within the reasoning worldview of the decision makers; an expansion and refinement of the existing 'natural heuristics' that decision makers use currently; a scoring and visualisation environment which is both fast and flexible but allows for, previously unavailable, levels of reasoning transparency and comparison. Taken in total the combination of the tool design, the heuristic artefacts within them and their influence on the hosts organisations, the two systems have proven they can provide an effective and valued "heuristic reasoning platform" for risks and issues. A future research direction is to explore ways in which the highly transferable heuristic artefacts in these systems, particularly measurement and data manipulation, might be strengthened via hybridisation with more powerful, but less transferred, formal systems like Bayes decision analysis.

Contents

Preface

This thesis is about risk. It is a contribution to answering the question: What are we to do about risk? I will suggest that the best thing we can do is harness it scientifically and socially to make it more than a dark force to be avoided, or an ephemeral goddess who may bestow fortune on the brave. This is not a new idea of course. I want to suggest however, that in the harnessing of it we have to understand it in some recognisable ways others will approve of.

Attempts to support reasoning about risk have not always done this very well in the past. This thesis is also about reasoning support. Existing theorists have not always been able to prove that risk and decision concepts remain recognisable to the communities from whom understanding has been drawn, or on whom theory has been applied. If those communities no longer recognise "their risk" and "their decisions" when they see them reformulated in available approaches, then we have to question whether the "support" part of the decision support earns its name, or indeed pays its way.

I hope to demonstrate in this thesis that I have created risk and decision support tools which address this problem and which span, comfortably or otherwise, the dilemma of normative versus descriptive decision science. The platform in this instance is risk, but I want to go beyond yet further theorising about the nature of risk and decisions, useful as this might be, to expanding the understanding of the concept of heuristic reasoning and applying it to risk.

Heuristics have tended to be confined to a sub-branch of descriptive approaches to reasoning. They are a kind of test case. The test shows where, or why, people get reasoning wrong and hypothesises ways to fix the bias this introduces. I want to come out in favour of an entirely new form of heuristic reasoning which nonetheless remains firmly within the etymology of that term. These I will call "essential natural heuristics" of decision makers. I will argue that

it is possible in applied researching to formulate these into both a bridge and a conduit to better reasoning for real time and distributed industrial decision making.

The "bridge" will take us to a better understanding of how industrial decision makers see the world of risk and decisions as communicable entities in strictly hierarchical and non-expert (in terms of risk) worlds. The "conduit" will allow decision support to create demonstrable increases in rationality. These increases I can prove were not only accepted by the communities as being about "their risk" and "their decisions" but the benefits these brought were valued and used.

As well as the essential natural heuristics of industrial decision makers I will argue that there is another group of "general natural heuristics" in evidence in these settings. These soft approaches, for example the use of "traffic lights", are widely considered to be satisfactory methods of reasoning to the groups I have observed. They tend, as I have seen, to be unquestioned and therefore generalised across decision types. These clearly inform and interact with the way people develop their own heuristics for reasoning. This has meant that strengthening the rationality of the general natural heuristics has been a key route to improving specific reasoning found in decision makers' essential natural heuristics.

I will argue that the closest theoretical school to my approach is that of Bayesian decision modelling. This in and of itself cannot be surprising since this school sits firmly between normative and descriptive decision science and, as has been argued elsewhere, is itself a form of "prescriptive" decision science.

Where I see the strongest parallel is in contending that Subjective Bayesian methods do not just "translate decisions into maths", but create a transition process for the decision maker to better reason about "self and world". This is a process that goes well beyond optimising

strategies based on expected utility. It goes so far beyond it that I could argue that the expected utility output is dwarfed by more useful organisational outputs caused by its strict pursuit. These might readily be of greater interest than they currently are.

Where I perceive divergence between myself and Bayes is in terms of my "technology acceptance" argument. The decision maker, even with the softer Bayes methods, has to take their reasoning out of their own natural reference grammar and accept a new mathematical "technology" to describe it. In comparing my work with prescriptive (subjective) Bayes methods I will show that I achieve a loyalty to the indigenous reference grammar and provide the decision makers with a "technology" and tools which embody it.

My approach could be described as a "near Bayes" or "pre-Bayes" modelling form. I will demonstrate that I have addressed the fundamentals of such modelling in common with Bayes e.g. eliciting attributes, utilities, combination and comparison rules and so on. I've done this to create a risk decision support which is quite formal but importantly, from the perspective of saying something new in this area, remains grounded within the reasoning worldview of the decision makers, their expertise and their own rationality. This I call 'for experts' systems.

I will argue, using two very different forms of risk as working examples, that moving outward from the essential natural heuristics of formal and distributed groups of decision makers is an equally valid path to decision support that ought to be considered and further formalised. My work shows this path to have yielded high levels of user acceptance. It may contribute to understanding and overcoming barriers to the use of powerful formal decision aids in real-world settings. These steps forward, particularly in risk decision support acceptance, may be a fertile ground for yet further developments in elicitation and decision science.

Chapter 1: Literature Review on Decision Theory

Overview

This chapter will review a range of the literature concerned with the formal definition of decision theories. That definition will move along a continuum from some of the formal mathematical models to some of the more psychological. Approaches like Subjective Bayes which effectively recognise the need to combine the two will be explored in some detail. There is also a phenomenological end to this continuum found in so called "real world" decision making approaches. These will be briefly reviewed.

1.1 Summary of chapter 1

The mainstream roots of modern decision theory and the debates around "normative", "descriptive" and "prescriptive" approaches to human decision support are of particular interest. The notion of whether we have the right frame of reference to study human decision making and so called error patterns therein is debated.

Normative decision theory has a "pure" statistical focus to create the rules and algorithms which exemplify normative axioms. Descriptive and prescriptive approaches introduce an applied psychological focus. This focus recognises that the satisfaction of normative axioms may need to compete with and/or be reconciled with observable decision-maker phenomena.

The question of biases and heuristics will be addressed, in particular the limits of currently accepted definitions in the literature. The key tension in this science, whether laboratory-sourced observations of well-documented heuristics generalise to decision makers in a real-life setting, will be discussed. Newer evidence around the persistence of the accepted decision making biases when considered under different framing conditions will be evaluated.

The subjective Bayes modelling paradigm will be considered in some detail for two main reasons. First, subjective Bayes can be argued to be the point in the above continuum where normative models and subjective agents meet in search of a real world model of a decision space. Second, it is one of the key assertions of this thesis that the applied reasoning models put forward in this research are a potential development of this paradigm.

Research situated at the social end of the scientific spectrum which goes beyond the introduction of subjective probabilities (considered contentious to some commentators on subjective Bayes approaches) is considered. These are seen as taking the argument for a continuum to a logical conclusion. These approaches can be seen to be highly task descriptive and focussed on the human decision making process. Whether the evidence for the efficacy of so called 'real world' decision support models is convincing remains in debate.

Main conclusions

Normative statistics currently remains the dominant frame in decision science. The discipline of Psychology increasingly contributes to the establishment of theories of decision making but this seems to focus on empirical experimentation. Some theorists discussed argue that this experimental approach introduces compromise by setting (necessarily) un-real test conditions.

The field of heuristics and biases is an example of the influence of experimentally based thinking, being itself a kind of blend of statistics and empirical psychology. The narrow view this field produced (that human reasoning is highly faulty) has become increasingly challenged on philosophical and methodological grounds.

There is a visible debate on "real world rationality" and the dominance of current statistical and empirical psychology frames of reference. Whether these frames provide an appropriate

sort of benchmark for reasoning in a real world setting is increasingly contested. Subjective Bayesian decision analysis could be considered a half way house in such a debate. It aims to fit normative models to real world decisions. What remains uncertain however, is whether the approach is entirely successful. That uncertainty hangs on the inherent complexity of its techniques. Even the so called real world decision theories can also be shown to be methodologically very top heavy for real world application.

In common with the direction of some newer statistical research this thesis requires a decision support methodology aimed at 'real time', 'distributed', 'high volume' and small-scale decision making tasks. This review concludes that this space would best be filled by a variant of subjective Bayes methods. The variation suggested exposes the potential efficacy of a new conceptualisation of heuristics in applied reasoning.

1.1.1 Digest of key points made in this chapter

- Statistics and psychology together should be the dominant frame of reference to understand decision support about risk. These should provide a dual notion of "rational" decision making
- Much extant work in each discipline is criticised for a failure to predict actual decision maker choices in real world conditions
- In statistics "the normative ideal" can only apply to non-real settings
- Heuristics and their effect on reasoning is shown not to be as well defined or understood as was previously believed
- The accepted view of heuristics and biases research that human beings have faulty reasoning may itself be biased by a research premise which expects participants to be fluent in probabilistic representations
- Later evidence from the same disciplines suggests the phrasing of problems in the original research phrased can account for some of the faulty reasoning
- The most appropriate framework for studying human reasoning may not yet be available
- Effective decision making using a subjective Bayes approach has many requisites for modelling which deliver a very large analytical burden
- The tendency of the extant applied research to consider problem spaces which are very large, and amenable therefore to significant investment in complex modelling, is considered a weakness
- The ideal features of a decision support system can however be synthesised

1.1.3 The problem of defining decision support

The term decision analysis seems to have a fairly firm set of definitions and expectations around it. Goodwin and Wright suggest that:

"Decision analysis therefore involves the decomposition of a decision problem into a set of smaller (and, hopefully, easier to handle) problems. After each smaller problem has been dealt with separately, decision analysis provides a formal mechanism for integrating the results. . ." (Goodwin and Wright, 2008. p3)

Having begun with this definition however Goodwin and Wright begin to outline how decision analysis does not, in most cases, do this:

"It should be stressed, however, that over the years the role of decision analysis has changed. No longer is it seen as a method for producing optimal solutions to decision problems." (p4)

The term decision support is perhaps less well defined. This may be because the highly technical decision literature, as we will see, is broken up into many specialist fields which do not have enough in common to be drawn under a banner like this. When attempts to define decision support are specifically made, because it is the support of decisions which is the field of interest, vague results can be seen, for example Baverstam has this to say:

What is decision-making support? Although this must be a question of great relevance here (radiation protection, my brackets), *it seems quite seldom asked. The answer is regarded as obvious and it is implied that it is the same to everyone. But is the answer the same to those who give support and those who receive it? . .The concept of 'decision support' includes much more than we usually think. . .motivation and confidence are of vital importance in relation to decision making support for emergency management"* (Baverstam, U., 1997, p1).

Searching for a better definition of this area than the four paragraph answer given by Baverstam results in the discovery of more long answers. The on-line encyclopaedia (http://en.wikipedia.org/wiki/Decision_support_system) has this to say:

Because there are many approaches to decision-making and because of the wide range of domains in which decisions are made, the concept of decision support system (DSS) is very broad. A DSS can take many different forms. In general, we can say that a DSS is a computerized system used for supporting rather than automating decisions. A decision is a choice between alternatives based on estimates of the values of those alternatives. Supporting a decision means helping people working alone or in a group gather intelligence, generate alternatives and make choices. Supporting the choice making process involves supporting the estimation, the evaluation and/or the comparison of alternatives. In practice, references to DSS are usually references to computer applications that perform such a supporting role.

This article (available June 2008) does give a very informative definition, but it also then goes on to explain how the definition of this area is very polarised over a whole range of academic disciplines whose definitions may or may not agree with or compliment each other.

Decision support, as it pertains to the systems in this thesis, is perhaps best defined simply therefore as follows:

Decision support systems are rational tools, methodologies and software which help people make better, more consistent, more transparent and more rational decisions across a range of problem types.

I will adhere fairly closely to the terms in this definition throughout this work.

1.2 The breadth of theory

The subject of this thesis lies across the interface of decision support / decision analysis, human reasoning, applied occupational psychology and risk analysis. The breadth of the available literature, and the significant disagreements therein, mean that avoiding the trap of a

post-modern vortex is important. The landscape of the theory base can be described in the continuum summarised below.

Increasing subjectivity

Normative ideal = True expected utility and value
= One fixed value outcome

Game theories = Choice and preference rules
= Combination rules

Game rules = Using bands of uncertainty

Subjective probabilities = Accepting these may not match reality
= Working with the best available model

All still trying to stay within the normative axioms

Increasing phenomenology

Descriptive ideal = Observation and analysis of actual behaviour to explain how decisions are made

Heuristics and biases = Deviation from normative logic

Judgement experiments and stylised hypotheses

Behaviour under games of chance

Still attempting to accept the world of normative logic and describing people as deviating from this

Increasing realism

Prescriptive ideal = Guide the decision maker in the evolution of the decision
= Sill using the axioms of probability
= Assisted modelling of the decision space
= Subjective but controlled reasoning

Semantics = Back to the real world
= Dilution of logical problems because of real world starting point

Blending descriptive and normative approaches

The correct frame of reference for the theory

Decision support is a statistical discipline. This is because there are a range of larger statistical themes (which remain fluid to this day) in areas such as axiomatic probability theory, subjective probability, expected value, expected utility, biases and paradoxes and the application of various functions and decision rules. These attempt to explain/assist reasoning process and produce/predict best outcomes.

Decision support is also a psychological discipline. This is not least because without decision makers (notwithstanding the idea of artificial intelligence) there are no decisions. As O'Hagan et al point out in the case of eliciting expert judgements (O'Hagan et al, 2006):

"the processes at play within elicitation fall fair and square within the remit of Psychology". (p33)

We might say that the frame of decision science is a kind of anthrocentric mathematics. This is because linked to, or perhaps reacting to, the complex statistical dimensions, there are a number of broad social and contextual themes. These include multi-criteria approaches (e.g. Stirling, 2001), naturalistic decision making (e.g. Beach, 1997) and heuristics and biases (e.g Plous, 1993). Presiding to some extent as a meta-filter over these is cognitive psychology (e.g. Oaksford and Chater, 2002). The applied psychologies would claim to deal with perceived shortcomings of formal decision theory to predict actual decision makers choices.

To build an adequate decision support system one has to address the efficacy of these themes. Statistical theories are useful because they comply with explicit boundaries and conditions. Those theories are considered "normative" science and are therefore highly valued in scientific circles and in quasi-scientific forum such as government bodies. Psychological theories are, unsurprisingly, more social and behavioural these tend to be collected into an overall idea of "descriptive" science. This is valued also. In decision theory the normative –

descriptive continuum is however the subject of a number of tensions not the least of which is the rise of the idea of a third, intermediate, "prescriptive" form (e.g. French, 1985).

Normative decision theory

Normative decision theory is not really used as a theory, rather it's analyses are set in a utopian paradigm from which current decision methodologies can be judged. Models which score well in this judgement are at home in the controlled conditions of the laboratory. Normative axioms are also extensively used in various game theories. In these theories all players are assumed rational i.e. satisfying the normative ideal.

The explanations of human rationality given by a normative approach are based on idealised decision pathways operating often on perfect information and using sophisticated maths. Decision theories in this school cluster around the concept of maximising expected utility of a decision. That utility can be a fixed value outcome e.g. financial reward, or it can be a strategic value outcome e.g. gains in competitiveness.

The normative decision maker can build a utility function to operate on key decision variables using known probabilities and generating an exclusive expected utility for each outcome. This utility provides the basis of a rational decision and should therefore logically drive (the logical decision maker's) preference or indifference (e.g. Oliver and Smith, 1990).

Challenges to a normative approach

There are a number of challenges in taking the idealised world of the normative science outside into the manifestly non-ideal worlds of applied decision making. Not the least of these challenges is the necessity, in normative reasoning, to have a fixed, unitary decision as the end point. Industrial decision making, particularly when it concerns human behaviours

for safety and business competitive variables, doesn't typically benefit from this kind of fixed or closed system. Such decision making happens under conditions of constant change and uncertainty.

In decision making under uncertainty a normative decision maker would be at a real world disadvantage. The outcomes, and their probability distributions, are not fully known, and so cannot have these explicit formal relationships. For example expected utilities cannot express in this setting mutually exclusive outcomes. Real world decisions have to be made without these levels of certainty and when that happens the normative approach has to be strengthened.

To strengthen the normative approach one can introduce game theory based rules. The 'max-min' rule is an example of this where the worst possible consequence of the chosen alternative is better than (or equal to) the best possible consequence of any other alternative. More complex Schaffer-Demptster type rules are another route (Schaffer, 1990) where, simply put, one still uses probability but creates intervals of uncertainty rather than a (normative) fixed score.

As an alternative solution one could attempt to elicit some *stable* form of subjective probabilities and use these as if they were normative. This solution comes with the (rather strained) caveat that these probabilities will still be admissible within the normative framework, an idea that remains contentious (Cooke, 1991).

Descriptive decision theory

A move away from the axiomatic model world of normative theorem e.g. by admitting subjective probability, is a move toward "descriptive theory". Descriptive theory is less

concerned with laws and rules and tries to describe, by various forms of observation and analysis, how decision makers are actually making decisions. Describing decision making and its processes in this way led to a school of thought on how and why we, often systematically, go wrong in decisions. A range of cognitive biases were first hypothesised from this method by Tversky and Kahneman (e.g. Tversky and kahneman, 1974, 1981, 1988).

In the complex world of gaming theory, attempts were made to accommodate these "errors", for example Larkey and Kadane developed a hybrid gaming theory (Larkey and Kadane, 1982, 1983) where one's subjective beliefs about an opponent could be expressed descriptively whilst one's beliefs about one's own reasoning could conform to normative Bayes methods. This is an example of an interesting tension to re-affirm the applicability of objective (normative) rules to an increasingly subjective subject matter and keep it within the normative sphere.

Descriptive methods studied judgements on stylised hypotheses or in games of chance in the fixed laboratory experiment. This approach might be considered to be "closer" to the real world, but clearly it is still not the real world. Normative science (both in decision making axioms and in the research methods chosen) still studied decision makers in "un-real" settings. This is perhaps just good stepwise science.

Prescriptive decision theory?

Early descriptive methods, especially those focussing on errors in reasoning, seemed motivated to develop "adjustment factors" which could re-align faulty human decision makers to a more normative model. If this has been more of a success, normative models could have gained ecological validity in controlled, but applied, decision making environments. Prescriptive approaches arrived perhaps because there was little evidence that this would be

the case. Still trying to utilise normative models, prescriptive methods aimed (more descriptively) to guide the "evolution" of decision makers' perceptions in the direction of agreed ideal consistency. This agreed consistency would of course never be a formal satisfaction of normative axioms. However, the normative appetite for complete closed order was resisted, as was the descriptive appetite for focussing on errors. Prescriptive techniques therefore attempted to moderate the limitations of the decision makers actual cognitive processes whilst still capitalising on normative claims (Bell, 1988).

This is a mark of the influence of normative science on the area and its practitioners. Descriptive methods were still holding onto normative logic and use it as a benchmark for their conclusions. Likewise prescriptive methods, whilst moving further along the spectrum were still placing a high value on the normative benchmark.

The challenge of the applied perspective

So far, staying within the mainstream classification of decision theory, we have three intersecting positions: normative, descriptive and prescriptive. The point of intersection is a desire to hold on to those powerful normative axioms and apply these into a form of "real world" decision space. A challenge arises if the decision makers in question are:

a. not in laboratories considering closed rational systems

b. concerned with the object of a decision not the subject of decision making

c. unlikely to want to become mathematically adroit in order to make decisions

One might argue a nuclear scientist, an industrial process engineer and an investment modeller would all fit well into (a) to (c) above professionally. That explains perhaps the prevalence of these sorts of disciplines in the applied theory. However, I would argue that the bulk of people who make decisions in industry are managers. Furthermore those managers

will not fit into (a) to (c) at all. Some of them, as we will see, make decisions which are causally linked to the daily safety (from life terminating events) of hundreds of people. Even in the case where things are not so dramatic, some of them will be managing situations which run to the hundreds of millions of euros, without themselves having any financial responsibilities or control.

This first sweep of the general theory already raises important questions about what sort of validated systems such managers could seek to use for decision support and increased rationality?

- If this field is as complicated and unresolved, where do real life systems begin?
- Should a real life system increase formal rationality e.g. by accepting the need to decrease known biases?
- Should system decisions be represented by formal logical rules, or should their idealised consistency be in a negotiated tension with their real-world application?

1.3 Biases and heuristics, the debates

In this next section

I will look at the question of biases and the potential framing of real world decision making from the heuristics argument. I will examine some of the growing tension around the applicability of laboratory-sourced observations of well-known heuristics to a generalised human decision maker. I will also look briefly at the evidence for the persistence of the recognised decision making biases when decisions are framed more naturalistically and phenomenologically.

I will conclude that there is a need to understand what sort of benchmark can be used for the notion of "rationality" in the real world decisions under question in this study.

Biases, rationality and real-world decisions

Should applied decision support decrease known biases? Whether or not heuristics and biases are seen as naturally descriptive or as an improvement to the usability of normative theory the challenge must be to understand them in live application. Take a trivial example (nonetheless a key one for one of the case studies): a consumer choice.

A consumer choosing a product on the shelf, be it satisfying a primary drive (like a food) or a socially constructed drive (like a de-odorant) will make that choice heuristically. It is argued that this heuristic choice will be a quick and adaptive one and only based on some of the available data. Although no elaborate calculated probabilities are used the choice does simplify an extremely complex environment (Feidler & Schmid, 1995) without the need or systematic and exhaustive analysis of all he available data. When heuristic processing is used to describe behaviour of this kind, rather than errors in stylised experiments, it can be argued to be a symptom of intelligence.

Move out of the supermarket and into the more engineering/scientific world air traffic control tower however and scientists seem less convinced that heuristic decisions are intelligent with respect to safety[1]. An indication of this is the degree to which the famous study of this area the biases project (Tversky & Kahneman, 1973) is used. This oft-quoted work is, of course, backed up by a considerable corpus of resultant study (their own and others).

Tversky and Kahneman were among the first people to propose formal kinds of human heuristic with stylised rules. Three heuristics are most noted:

- The availability heuristic refers to the tendency for an event to be rated more probable to the extent that it is more easily pictured or recalled. E.g. people over-estimate the

[1] This is not to underplay the largely invisible, but considerable, decision making in product safety science

risk to life posed by 'sensational' events and underestimate of the risks associated with mundane events (Lichtenstein et al, 1978).

- Representativeness heuristic is when people judge probabilities "by the degree to which A is representative of B".
- Anchor and adjustment is the tendency for the decision maker to anchor on their own or a "first guess" at, or suggestions for, the answer to an uncertain problem and then adjust to that, as opposed to rejecting it completely.

Were we to apply evidence of these to air traffic decision support we would be shaping a decision support system in a particular direction. However, it may not be too wise to do that. There is a profound (and growing?) tension in this area of science. The work of Tversky and Kahneman remains (on a surface reading) the most oft quoted, but newer research is calling it into question.

'Uncertain Judgements' (O'Hagan et al, 2006) is a multi-author book looking at eliciting experts' probabilities. Their review chapter on "The Psychology of Judgement Under Uncertainty" it thoughtfully questions the supremacy of Tversky and Kahneman's Prospect and other theories in human decision making. The book supports this by discussing newer successful research into a number of alternatives and counter positions.

The introduction to a section on the psychology of judgement under uncertainty traces the change in zeitgeist from Peterson and Beach's 1967 view of humans as good "intuitive statisticians" to Hogarth's "selective, stepwise" and "limited" human in 1975. The catalyst for the change was of course compellingly directed by Tversky and Kahneman's work on judgement under uncertainty. O'Hagan et al, 31 years into the debate, do not seem as convinced:

"All agree that the demonstration of judgement can be flawed does not imply that it always will be flawed. Nevertheless there has been considerable difference of opinion expressed over just how good or bad judgement under uncertainty is. . " (P36).

Tversky and Kahneman were interested in the instinctual processes people use to make judgements. Their most famous finding (which O'Hagan et al also call "oft cited") are certain proposed heuristics: availability (the tendency to associate higher probabilities and larger class sizes to things that come easily to mind), representativeness (judgements made according to similarity between instances) and the anchor and adjustment (the tendency to only make small step changes from your own, or other's initial quantifications). Their proposition being the first, and, given the state of the art at the time, the most compelling evidence, set the mould for the use of this term into quite a narrow band. That band was of essentially studying human error. Unlike Rasmussen's (Rasmussen, 1982) conceptualisation of error which was to look at known cognitive processes in real world task settings, Tversky and Kahneman were perhaps filtering theirs through unchallenged normative models of probabilistic judgement. That in itself made fluency with the notoriously difficult area of probability theories a prerequisite for criticism.

O'Hagan et al come well equipped in that department. Having laid down a feeling that they may challenge these heuristics they catalogue a number of mechanisms in support of the availability heuristic and in support of broad anchoring effects in line with those highlighted by Tversky and Kahneman. Looking at representativeness however yields a more interesting debate. Rather than just accept "the conjunction fallacy" (where the apparently representative nature of data over-writes the logical use of probability) as an example of this bias, Fisk's (2004) alternative accounts are cited. Likewise in "base rate neglect" (e.g. where people ignore longer run average frequency in favour of frequency under review) also purported to

support representativeness bias, O'Hagan et al begin to consider other explanations, the cornerstone of which is a semantic solution.

Semantic theories criticise the way in which humans are expected to operate in these experiments e.g. being given obscure, or unfamiliar, probabilities rather than, for example, transparent frequencies (which lessen some effects). A semantic approach questions the legitimacy of the framing of the solutions in purely mathematical terms (Cohen, 1981). In the "confusion of the inverse" (where people act as if an inverse probability has an interchangeable value with a stated probability) confusion resulting from the peculiar semantics required to discuss probability could be at the heart of a "poor conceptual grasp" (Koehler, 1996) of conditional probabilities. This is rather than being a consistent bias. Semantic approaches have the makings of a critique of the experimenters own framing of the explanation (Meyer and Booker, 1991). O'Hagan et al begin to look for evidence that:

"perhaps errors in judgement. . .occur only because the questions that elicit them are phrased in such a way as to encourage biased reasoning."(p46)

Once detailed discussion deepens so does the authors' challenge to heuristics. Fieldler overturns the famous "Linda is a feminist bank-teller" conjunction fallacy example by using frequencies rather than probabilities reducing Tversky and Kahneman's 50% rates of error to 20%. Adherence to Bayes rules in the "confusion of the inverse" problems likewise reduce when frequencies are used (Gigerenzer, 1996). The magnitude of the "base rate neglect" also dropped in the frequencies condition (Griffin and Buehler, 1999).

Jones showed that frequency formats have a complex interaction with the heuristics. The availability heuristic is essentially a relative frequency paradigm and the representativeness, essentially a single case paradigm (Jones, 1995). In the latter case bias is reduced by

introducing frequencies and in the former it is actually increased. "Cueing effects" are prominent particularly those to do with the phrasing of questions in reasoning experiments (Schwartz and Strack, 1991).

From a research methods point of view, a final criticism is levelled at Tvesky, Kahneman and others, perhaps "the lost assumption" of laboratory-based decision science. The tendency to conduct research on available populations of students creating an important (and rarely mentioned) sampling bias (Sears, 1996).

In the face of criticism for heuristics?

The heuristics and biases approach is, in and of itself, being challenged in many quarters now. The challenge comes close to suggesting that too rigid adherence to these explanations of behaviour is an example of a heuristic itself. Likewise the tendency to fit research design and findings only into this prevalent paradigm is starting to look like a bias. If a decision support system wanted to reduce bias there might be benefit in understanding some of the possible sources here and assessing their relevance. With this in mind, our second question becomes important: should the decisions be represented by formal logical rules, or should their idealised consistency be in a negotiated tension with their real-world application?

Support Theory and how to describe decision spaces

Support Theory is an example of a statistical theory, using a normative approach, which tries to addresses this. Support theory, is a mid 90's to the present day attempt (covered by O'Hagan et al), to try and gather up the psychology of the way that people assess probabilities (a key to unlocking any risk decision). O'Hagan has, in two of the famous heuristics, shown claims (from support theory) that:

"different description of the same event can give rise to different probability judgements".

This is a problem for Normative schools of decision support which would assume "description invariance". Support theory suggests there exists an "unpacking principle" which skews the ability of the individual to judge probability. For example, a more detailed hypothesis for a future event will earn a higher probability than a simple one and judgements for unique and exclusive events can exceed probabilities of one.

Support theory shows that heuristics and biases can in fact be reversed by causing the individual to focus on, and admit, more data. The data 'unpacked' in this way within the normative style probability statements ruins the logic of the probabilities people use. They are somehow lured into reading more probability into the situation than can logically be there.

What kind of rational?

What this demonstrates is there is a body of scholarly research which successfully undermines a simple acceptance of a Tversky and Kahneman's based (or indeed 'style') biases approach to human reasoning. The early "definitive" work on heuristics and biases is now roundly challenged for its applicability to the sphere of decision support in question in this thesis. That challenge however, still comes from within the very traditions that generated the theory that is now being criticised. Therefore many of the framing assumptions surrounding decision makers, and their foci, remain i.e. a definition of rationality which is still normative in character and a definition of heuristics, with very few exceptions, which is still a faulty attempt at normative rationality. For a thesis like this is an important factor in the choice of decision support design. There is not only a lot of disagreement in this area, but there is still only really one area.

1.4 Some Psychology of reasoning

In this next section

In this next section I will review some of the more highly detailed work that has been built up about reasoning within the cognitive psychology paradigm. This is going to cover a broadening of the idea of heuristic reasoning looking briefly at comparisons with computer modelling. This will also show a deepening of the debate about whether heuristics are highly limiting of reasoning, very powerful aids to reasoning, or indeed whether we do not yet know enough to say.

I will conclude that these studies illustrate how the area is highly academic and has potentially become about a nuanced debate around the philosophy of science.

Cognitive approach?

The significant recent work of psychologists Oaksford and Chater has done a great deal to further broaden the debate not just on heuristics, and their legitimacy in normative methods, but on the assumptions base of human reasoning research in toto. In so doing they challenge core principles which guide much of the research effort already reviewed. Their work opens up an important debate on the selection of a decision support method particularly because they suggest that the most appropriate framework to study human reasoning is, in fact, not yet available. In their book: Commonsense Reasoning, Logic and Human Rationality (Oaksford and Chater, 2002) they state:

"Although many theorists have argued that deduction is at the core of cognition, we argue that it is at the periphery. From this point of view we shall argue that the "errors and biases" observed in "deductive" tasks in psychological experiments should be understood not as failed deductive reasoning but as successful non-deductive reasoning. Consequently these

"biases" do not provide evidence for human irrationality; rather the reveal the nature of people's commonsense reasoning strategies." (P 179)

In common with O'Hagan et al Oaksford and Chater suggest that the field of heuristics is not as well defined or understood as has been believed. An example of this would be the debate in Chater's peer review comments on Gigerenzer, Todd's (and the ABC group's) thesis on "simple and fast heuristics" (Chater, 1999). I will expand on this example (a critique of the 'Take the best' heuristic) below but the argument between the theoretical peers is, in and of itself, very informative.

Gigernezer and Todd seem at once to be praising and highlighting deficiencies in human reasoning with their "simple, fast and frugal heuristics" approach. On the one hand they suggest that heuristics people use to reason are actually very smart and efficient. On the other hand they note that a computer algorithm to replicate some of this reasoning is rather simplistic. The heuristics in question, and by implication human cognition itself, may therefore not be all that impressively complex after all (notwithstanding exceptions like language recognition).

Chater's counter to this is that it ignores the evidence that in many other areas humans vastly out-perform computers. He criticizes Gigerenzer and Todd for tending to focus in (deliberately?) on those laboratory tasks where humans are exceptionally weak to prove their heuristics thesis. Such a focus is considered unacceptable when it is to the detriment of research into understanding the great many tasks where humans have hugely powerful reasoning capability. Unfortunately this is research which is, lamentably, not available.

When reviewing a range of current theory from Epistemology, Psychology of reasoning and from computer science Oaksford and Chater are forced to conclude:

"Nonetheless, only further research within each of these three approaches will provide an answer to the question of which is the most appropriate framework for studying human reasoning". (Oaksford and Chater, 2002, p 206)

Is there a confusion between school and theory?

This "fast and frugal" rationality argument pertains to have come from the "ecological" school of rationality. It says, if basic algorithms, which do not conform to classical ('normative') decision rules, can match, or out-perform their 'normative' equivalents then why are these equivalents given priority? A key example of this is the "Take the best" algorithm. This algorithm is not supported by rational norms, but, importantly, it has been shown to be as successful as linear regression methods in judgement tasks (Gigerenzer and Golstein,1996). Not only is it a success but it uses less data, applies fewer rules and takes less time. Gigerenzer argues that these and other observations show that normative theory should not be the benchmark for human reasoning.

Chater et al have argued very much the opposite from the same data (Chater et al, 2003):

". . . that norms of classical rationality are crucially involved in explaining why a particular behaviour is ecologically successful. Thus, we argue that classical and ecological notions of rationality are complementary, rather than standing in competition." (p65)

This position suggests "the standard notion of rational explanation" does in fact, contrary to Gigerenzer's view, refer to a rationality which both stresses rather than ignores the environment and takes cognitive limitations into account. They also provide evidence that *other* algorithms, from within the classical model, perform equally well compared with Take

the best. Thus they conclude that Gigerenzer's theory has *"run ahead of rational explanation,. . ."*(p82). The error has been to view *". . . the human mind only as a probabilistic or statistical calculating machine."* (p82)

This argument is over the etymology of rationality itself where Chater et al conclude we must:

". . . emphasise the importance of the project of providing rational descriptive explanations that can explain when and why the cognitive system is adaptively successful." (p82)

Oaksford and Chater suggest work on the probabilistic approach is heading in this direction (Oaksford and Chater, 2001). Again, in defence of a better formulation of reasoning with respect to biases and heuristics, they argue that reasoning failures can be understood from within a normative model. These are not based therefore on too selective a normative model for comparison, but on the way people make sense of the everyday world:

"A probabilistic approach to these processes explains people's performance in the laboratory as a rational attempt to make sense of the tasks they are set, by applying strategies adapted for coping with the uncertainty of the everyday world. It is these strategies that create the appearance of biased and irrational reasoning when compared with the standard provided by formal logic." (p.356)

Are reasoning theories adequate?

In considering the debate above one can own up to a certain confusion. The trouble is this, theorists appear to swap places in these arguments. The normative and descriptive landscapes seem to be two and a half sides of the same coin. This argument is still taking place at the intersect between normative, descriptive and prescriptive decision theory although perhaps now we have to add ecological theory onto the continuum. In the introduction to the same

paper Evans and Over's dichotomy between "rationality 1" the sense in which people perform sensibly in achieving goals and "rationality 2" the kind provided by normative theory, is rightly described (Evans & Over, 1997, p. 1) but not convincingly applied.

This sort of partial disagreement confusion was highlighted in the work of Evans in his four way (Evans, 1991) classification for reasoning theories:

1. The mental-logic approach with formal inference rules (e.g. Inhelder and Piaget, 1958).

2. Mental models theory, which emphasises the power of semantics to create more powerful representations in the cognitive system (e.g. Johnson-Laird and Byrne, 1991).

3. Pragmatic reasoning schema theory, proposing content specific inference rules (e.g. Cheng and Holyoak 1985).

4. The Heuristic approach proposing systematic errors and biases due to cognitive short cuts (e.g. Evans himself 1989).

Concerned that all of these theories offer "cognitive limitation" as a central tenet i.e. that the human brain cannot adequately process large or complex amounts of information in real time, Evans tried to get these theoretical schools to reach a consensus. This consensus was to be over how we in the outside world might judge the adequacy of these theories to:

1. Explain human competence, the fact that human beings solve deductive problems all the time.
2. Explain formal biases, how is it that many errors which are made are systematic i.e. not the result of random overload but somehow in a pattern.
3. Explain content – context effects, why is it a fact that the same logical problem with different content or context can radically alter people's ability to solve it.

Evans' view was that the four main theoretical schools had founded their theories upon only answering one of these three questions. He appealed to the philosophy of science's historical standards for theory i.e. completeness, coherence, falsifiability and parsimony.

Oaksford and Chater having described Evans' work (Oasksford and Chater, 2001) argue, over and above this, that all theories which imply cognitive limitations approaches to people's rationality fail to recognise the key inferential mode found in humans. That mode is of selecting appropriate information from long term memory for reasoning. Only Artificial Intelligence (AI) has broached this area, and then with a great deal of difficulty.

Oaksford and Chater propose that the whole theory of reasoning that surrounds these famed demonstrable human cognitive limitations is for artificially limited laboratory tasks. These tasks are selected to the detriment of others which would be able to begin to explain the awesome power of human cognition to store and call upon a life-time's worth of long term memories and to reason from them. Most mainstream theory of reasoning might therefore be insufficiently generalisable to real humans. Chater notes that even in artificial intelligence the problems of "scaling up" support this. Simply put, AI programmes which can handle "toy" problems rather well, when using the database of information that would be required by a laboratory experiment, failed when scaled up to deal with the more realistic data sets facing every human every day. (e.g. McDermott, 1987).

The nub of the matter seems to be a nuance of normative theory surrounding a focus on deduction and inference. Chater, Gigerenzer and others would seem to agree that since very little of human reasoning behaviour actually involves deductive reasoning, why should normative laboratory experiments focussed almost entirely on deductive tasks *of logic* be the bed of theory? Where the disagreement comes in seems to be that Chater and Oaksford wish

to maintain the normative proposition but steer it into other research, whereas theorists like Gigernezer and Todd wish to reject it in favour of new-wave heuristics.

Taking stock from an applied perspective

At the beginning of this section I asked should decisions be represented by formal logical rules, or should their idealised consistency be in a negotiated tension with their real-world application? This review of the work of some key authors has served to further open up that debate rather than answer the question. It has become a debate about the philosophy of science, the standards for evidence and the meaning of "reason". One school suggests the evidence supports the rejection of normative standards, another says the same evidence is actually in support of it, a third suggests that the theories are inadequately differentiated.

These decision support approaches discussed so far, which might underpin a future application of heuristics in real world risk-decision support, clearly show some of the tensions of a post-modern paradox. The normative science as "pure science" (i.e. applies to an unattainable model world) the descriptive science as a bridge which attempts to realise it in real world settings. The over estimation of either turns explanation it into "an art", which is why the debate proceeds primarily at the philosophical interface.

What I think we can say is clear in three areas:

1. The classical rules of normative theory, particularly with respect to probabilistic reasoning, may be judiciously applied.
2. A great deal of received wisdom on the "biased human" may itself be biased due to sampling, methodological and paradigmatic choices which seem to have placed it on the edge of being self-fulfilling.

3. Human beings have a huge power to reason in real world settings and these should shape the context of our understanding and support of decision making.

1.5 Observations on real world decision making

In this next section

There is a raft of research situated at the social end of the spectrum which is concerned not as much with logical reasoning forms as with reasoning efficacy in real world systems. I shall take a brief tour of this work.

I will conclude however that, despite the best efforts of the field, a highly transferable and uniquely heuristic approach to the real world risk reasoning in question may not be easily found. These theories are methodologically very top heavy and all give serious problems of transferring these technique into a real time distributed system for the high volume small scale risk reasoning in question.

Reasoning power in the real world

Humans have huge power to reason. Many of the authors reviewed so far agree on the importance of the real world context to release this. There are a number of formal decision approaches which frequently cite their "real world" credentials. We will consider briefly: Naturalistic Decision Making; multi-criteria approaches; deliberative mapping; multi-attribute utility analysis and value based thinking.

Naturalistic decision making

The term "naturalistic decision making" (e.g. Klein et al, 1999) is used to refer to decision making as it occurs in certain classes of real-world situations rather than the rarefied experimental conditions or highly controlled studies. It is specially studied in tasks such as life saving, critical medical decision making and so on. These are typified by high risk/consequences, ill structured problem spaces and stressful conditions. What Klein and others have consistently been able to observe is that highly experienced decision makers tend

to make effective decisions under these conditions. Note how this runs counter to the idea that high risk settings require more formal methods due to the risks being managed. Furthermore, this performance does not translate to experiments with normative style conditions and data. This lends further support to the idea that people function very highly in the contexts where human rationality actually matters.

Multi-criteria approaches

Multi-criteria mapping approaches are a decision analysis technique which deals with perceived shortfalls in conventional decision analysis. For example the assumption that all relevant data is available, the consequences of actions are knowable and so on. Mapping approaches which are multi-criteria tend to hold in tension uncertainty and plurality in the search for alternatives. This is seen as particularly useful where there is no uniquely rational way to resolve contradictory perspectives (e.g. the sitting of waste incinerators). A key aim is to interact with the decision space, broadening the values base of the decision away from a focus on technical and scientific data and information. Theorists would argue that the technique *"more realistically reflects the multi-dimensional nature of reality"* (Stirling, 2001). This lends further fuel, particularly when we come to discuss risk, to the idea of the real (scientific) reality and the perceived (or maybe that should be experienced) reality are considered different.

Deliberative mapping

Deliberative mapping', is an alternative form of multi-criteria mapping which brings together panels of citizens and specialists into joint discussions about how best to resolve highly complex problems. This complex process guides participants through a series of interviews and workshops, working together to evaluate a range of policy options. The deliberative process is: discussing a range of core options; developing a set of criteria or particular reasons

by which to judge the options; and then testing the options against the criteria; weighting the subjective importance of the criteria, providing a final ranking for each policy option where both criteria and weightings are recorded.

"Rather than provide a single answer, it allows many policy options to be considered at once. It reveals the key issues amid the complexity and shows who thinks what." (Stirling, 2000).

The downside of such processes is in being very demanding of time, resources and effort. It is usually only applied to large-scale, long-term, multi-stakeholder decisions on complex issues. It is noted however that an important by-product of processes like this is that they change the way the participants reason about the problem space.

Multi-attribute Utility Analysis

Multi-attribute Utility Analysis is used in health and environmental decision problems. Leaning towards more normative techniques it's strength is to evaluate options against multiple and competing objectives. It has five basic steps: identify objectives, establish attributes, quantify uncertainties, calculate strategies for maximising expected utility. Formal influence diagrams can be used to structure the approach more carefully. Challenges in this technique include the number of variables and maintaining the credibility and comprehensibility of analysis to diverse experts who contribute. This is still a descriptive technique but it has more rigour than the multi-criteria processes above. The advantages of this analysis is in its top-down nature and its focus on problem structure *"attention is focussed initially on key variables and their relationships, and only later on the details"* (Merkhoffer, 1990)

Value Based Thinking

Value based thinking, according to Ralph L Keeney (Keeney, 1992) suggests that:

"focussing early and deeply on values when facing difficult problems will lead to more desirable consequences. . ."

In a clean break with other methods, values are seen as more fundamental to a decision problem than alternatives. Value based thinking takes issue with the crop of current decision theory because their primary focus is alternatives. The "alternatives focussed" decision making techniques, wrongly for Keeney, figure out what alternatives there are and then support the choice of an the optimal one (from those which are available). Value focussed thinking consists of two activities, first deciding what you want and, second, figuring out how to get it. Keeney argues that this is not only more relevant but mirrors more accurately what real people really do. On the basis of some of the evidence in the previous section humans might be at the heights of their reasoning prowess here.

Taking stock from an applied perspective

On the face of it many of these methodologies would seem to be good candidates for a risk reasoning system. However their specialist nature is also their Achilles heel for a general transfer argument to small scale risk based decision support. Naturalistic decision making is clearly to experientially driven requiring real time risk encounters to function. Multi-criteria mapping and deliberative mapping are both purpose designed for multi-stakeholder environments involving internal and external players with multiple conflicting value systems. Mutli-attribute utility analysis is still too highly focussed on one large scale decision or strategy. Lastly, value based thinking is perhaps too philosophical. Whilst it might free applied decision makers from fixed frames and error prone reasoning, it will still return a highly naturalistic and aesthetic decision as an output.

1.6 Subjective Bayesian Decision Analysis

In this next section

I suggested at the outset that the subjective Bayes modelling paradigm was going to be the most likely candidate from which ideas might transfer into my applied problem. With this in mind I shall take a fairly detailed look at the area, hopefully confining this to the reference points we have already established in this review so far.

I will conclude at present that whether subjective Bayes decision analysis could be the model for the applied systems I need remains uncertain. This will be picked up again in my discussion. This argument will hang on ideas of complexity, acceptability, time and usefulness along with a blend of the ideas of focus and narrative.

What subjective Bayes approaches do appear to do is come near to (and sometimes exceed) my approach in terms of their organisational and rational sympathies. However, as will be seen, they are rendered 'top heavy' due to the possibility that there is an over-reliance on fitting the normative model to the real world.

What is very interesting about this approach is the way in which the need to reconcile objective and subjective criteria in a simplified structure is very apparent. It is as important at this stage as at any other to understand that these concepts are themselves a kind of short-hand for a far more detailed debate about probabilistic reasoning itself.

Subjective decision analysis

In the eyes of a social scientist the subjective vs objective debate may seem to refer to a simple dichotomy of science itself. Such an over simplification belies the underlying

intellectual debate. Bruno De Finetti states as one of his "preliminary clarifications" on the theory of probability:

". . . the important thing is not the difference in philosophical position on the subject of probability between 'objective' and 'subjective", but rather the resulting reversals of the roles and meanings of many concepts, and, above all, of what is 'rigorous', both logically and mathematically. It might seem paradoxical, but the fact is that the subjectivistic conception distinguishes itself precisely by a more rigorous respect for that which is really objective, and which it calls, therefore, objective. . .The subjective opinion, as something known by the individual under consideration, is, at least in this sense, something objective and can be a reasonable object of a rigorous study." (De Finetti, 1974, p6)

Interesting to note here is that De Finetti does not fall into the standard trap of assuming that subjective judgements are to be rated the inferior of so called objective ones. He is cognisant of the idea that users themselves may do this:

"You should be aware of superficiality. The danger is two-fold: on one hand, You might think that the choice, being subjective, and therefore arbitrary, does not require too much of an effort in pinpointing one particular value rather than a different one; on the other hand it might be thought that no mental effort is required, since it can be avoided by the mechanical application of some standardized procedure." (De Finetti, 1974, p.179)

Frank Lad, rather more metaphysically at times discusses the *"ill ease felt by subjectivist statisticians who attempt to engage in the professional arena of scientific statistical analysis, still dominated by directed activities of searching for true, unobservable, randomness generating structures.*

When challenged, the proponents of the subjectivist construction cannot solve the myriad well-known methodological problems of statistics that have arisen during the recent half

century of the dominating formalist-objectivist alliance. The reason is that form the subjectivist perspective, many of the proclaimed problems evaporate in their own metaphysical airs. We have no language with which to resolve them, as none is needed. The can merely be dismissed." (Lad, 1953)

Lad goes on to discuss an intellectual oppression of subjectivist approaches within which he suggests the victim is actually statistical modellers who *"have increasingly secluded themselves"* to the detriment of the believability of their science.

It is important as we come to discuss different subjective approaches that these three pitfalls, that of simplistic judgement of subjective data, that of allowing reasoning systems to provide this data for us be used and that of ignoring the lack of fit between objective science and the realities to which it pertains to help scrutinise what is being said.

Subjective Bayesian approaches

Bayesian analysis attempts to "utilise", or combine, the information in a decision problem which is a blend of scientific and experiential or judgemental data into the best course of action. To a true Bayesian the result is not just that course of action itself, but the ability to communicate the rationale for why it is believed to be so. This is the first key area where the thoughts in my system resonate with this approach. A Bayes decision maker will be able to: *"provide a framework within which is ideas can be critically appraised and modified, especially in the light of new information not originally incorporated into his model"* (Smith,1988).

This desire for a rationale to be generated along with a decision and the admissibility of new data is what fundamentally separates Bayesian analysis from normative models. Early applications of Bayes were still within a normative paradigm, Rational Bayes works using the theorem but with a set of utilities and probabilities. The paradigm soon widened. The

application of the theorem to belief nets and associated decision tree approaches have attempted a fusion of descriptive techniques with normative problems of expected utility. Subjective Bayesians would argue that rather than being a formal theory of decision analysis per se, it should provide an underpinning theoretical construct. This means that a number of different decision support activities now carry the name and their relative merits still divide the science (Cooke, 1991).

Goldstein and Wooff attempt to summarise "the Bayes linear approach" which is situated neatly at the crossover point between formal normative Bayes models and allowing more subjective judgements to enter an analysis. They state:

". . .we may view the Bayes linear approach either as offering a simple approximation to a full Bayes analysis, for problems where a full Bayes analysis would be too difficult or time consuming. . .or as a generalization of the full Bayes approach, where we lift the artificial constraint that we require full probabilistic prior specification before we may learn anything from data. (Goldstein & Wooff, 2008, p6.)

Goldstein and Wooff attempt to formalise their approximation of Bayes in a description with thirteen features. These emphasise a subjective starting point. Interestingly it is seen as important that the analyses have to be "within the ability of the individual to make". Expectation is substituted for probability in order to avoid the need to determine all probabilities in advance and allowing a focus on uncertainties allowing a "partial belief structure" to be constructed. Beliefs in this approach are fitted by linear fitting which is an approximation and computationally more simple, although this retains the structure of a Bayes model. The computational burden is therefore reduced. Over their systematic structuring of the approach what becomes evident is that the simplification is intended to " develop methods to assess whether "belief judgements appear intuitively reasonable."

Subjective Bayes decision analysis can be differentiated therefore as a modelling cycle, not the application of a model. As a whole process (both psychological and statistical) it is not a computational method alone. It works with ones beliefs about the probability of future under uncertain conditions. As such subjective Bayes modelling runs against classical statistics. Where the former attempts to interpret the evidence against a real world model, even if that is intuitive, the latter interprets the evidence through a series of laws independently of the world. A Bayes approach admits openly that a scientist interprets her evidence in the light of a world-view, in some circles this is seen as a serious bias of the method (Meyer and Booker, 1991). Classical statistics omits this possibility on the grounds that it cannot conform to statistical rules.

Subjective Bayesians would argue that a judgement based use of probability can escape its data by introducing updated, judgement-based uncertainty. To prevent a rationality loss Bayesian probability judgements, like those in normative models, are still linked back to axioms (often stylised here as betting preferences).

All so called subjective probabilities can be shown to suffer from biases under certain conditions (Hogarth, 1987, Kahnemann et al, 1983., Wright and Ayton, 1994) in this way they are very like normative / descriptive models suffering from biases of : availability; 'imaginabilty'; anchoring ; conservatism; base rate neglect; misconceptions of randomness; poor cognitive information processing; affective discomfort and so on. The subjective Bayesian would argue that scientific consistency is achievable in overcoming these biases via the judgement support procedures the technique puts in place.

Towards Prescriptive Analysis

"Most decision analysts now talk of prescriptive decision support and prescriptive decision analysis as being the application of normative ideas, mindful of the findings of descriptive decision studies, to guide real decision making." (French, unpublished notes 1998)

A good example of another analytical approach to fitting Bayes models to the needs of the decision support space what has been called 'the expert problem' (French, 1985). There are three classes of problem space, I have named them consultants, consensus group and bench-markers

1. Consultants: A group of experts is approached by a decision maker for advice (e.g. a share buyer consults a group of stock brokers). Here the decision maker is not the expert but must assimilate the expert judgements into a decision
2. Consensus group: A group who is responsible for a decision to the outside world (e.g. a series of managers assessing the safety of an air traffic control technology) who will look for a structured, rational, fair and democratic way to combine their judgements.
3. Bench-markers, a group of experts who may simply be setting a benchmark to be used in as yet un-revealed circumstances.

For 1 and 2 the focus is the decision and how to combine judgements. For 3 the focus is summarisation. In the combination of expert judgements (like risk and issues) calibration is therefore important. This calibration is done formally in Bayes elicitation and is even called de-biasing: comparing the experts predictions with the real outcomes (assuming one has them). A refinement to that technique relates to how informative (i.e. accurate) the experts judgements are in any given instance not just averaged over a long run (which might show high calibration value). This sorts of highly detailed thinking to provide support systems for possible irrationalities take us back through the descriptive norms and into a more fully developed idea of prescriptive analysis.

Prescriptive analyses use normative models to deliberately effect an evolution of decision makers' perceptions. It is one of the purposes of the analysis that the modelling process should be creative, dynamic and cyclic following a process of preference assessment and modelling, insights based on model exploration which leads to revised judgements which themselves lead to a revised model. This process is supposed to cycle until no new insights are found. This is a clear reaction to decision analyses based on a fixed or immutable view of preferences.

In attempting to build fully prescriptive systems there is something fundamentally sympathetic about the way prescriptive Bayesians set about supporting decision makers towards having a coherent model. French outlines a series of criteria in the choice of prescriptive methodology. In brief:

- Axiomatic basis: assumptions underpinning normative methods used need to be clear explicit and acceptable to decision makers and deciding behaviour reflects the ideals that decision makers aspire to
- Lack of counterexamples: the absence of any unsettling counterexamples
- Feasibility: the choice of practicable methods, i.e. not too many inputs, no over powered computations needed and no high dimensional representations
- Transparency: the users must understand what the inputs mean how the calculations have been performed and what the results mean. Not a black box.
- Robust: sensitivity of the analysis should be understood and a lack of arbitrary assumption
- Philosophically compatible: should fit with the decision makers view of the world and help them explore perspectives within that worldview.

"In short the acceptability of a prescriptive methodology will depend upon the philosophical outlook of the users."

If there remains a lack of decision making clarity, problem formulation tools can be used to reduce the lack of clarity (e.g. Soft OR methods Rosenhead, 1989).

On the use of subjective probability

Before we discuss some well understood problems with the use of subjective probability it is important to contextualise this further within the debate raised by De Finetti and others earlier. Whilst we are critiquing the use of subjective probability to inform decision making, that has to be seen as coming within an overall critique of probability itself. In the introduction to his book on the foundations of statistics Savage ruefully points out:

"It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. There must be dozens of different interpretations of probability defined by living authorities, and some authorities hold that several different interpretations may be useful, that is, that the concept of probability may have different meaningful senses in different contexts." (Savage, 1954, p3)

Nonetheless, there are some well recognised problems with using subjective probabilities. We have already mentioned phenomena such as conservatism, see Smith's version of the tennis match example (p45). Here, in the face of clear logic, individuals using probabilities under guidance, rather than as experts, still produce exclusive probability spaces which do not sum to one. The assumption is that subjective probabilities can be fitted to a classical model with relative ease, according to Smith this assumption has been *"criticized as being too strong"* but as he says *"it forms one of the cornerstones of Bayesian decision analysis"*.

There are a number of fixes the decision analyst can introduce to stiffen up subjective probabilities such as: Using betting schemes which are very similar to your clients problem; breaking down events into components; avoid probabilities which are either very large or very small; differentiation of joint probabilities.

The counterarguments from within subjective Bayes to the challenge that the use of subjective probability (or non axiomatic) is manifestly "illogical" seem to appeal to some kind of intangible not being picked up in the decision analysis, for example self esteem. They also to appeal to some kind of cognitive connectedness to the subject matter which doesn't happen when the subject matter is raw immaterial, such as tossing a coin. There is enough debate highlighted about this to suggest that it is at a bit of an impasse. This impasse raises questions about how much support (and how long lived) clients would really need. As Smith says:

"In practice the probabilities you elicit will invariably not satisfy the probability axioms if you happen to ask enough questions, though the client is usually prepared to adjust them when you point this out and treat the inconsistencies as measurement errors" (p47)

A question of process and content

As is clear from this detailed consideration, subjective Bayes modelling is attempting a thoroughness which goes beyond the application of a decision making theory and into the heart of what an organisation is trying to achieve by clarifying the decision space. Philips called this "requisite modelling" (Phillips, 1989), looking at large scale technological processes he proposed that "decision conferencing", an intensive two day problem solving exercise facing complex issues, was a form of "social decision analysis" and that adequately conducted these produced "requisite models". A facilitated process the aim is to arrive at a

simplified, though not simplistic, structural form for representing the problem. This is used to house both firm data and subjective judgements about the problem.

Phillips concludes that decision conferencing:

". . . lends structure to thinking, and allows all perspectives on a problem to be represented and discussed. . .the process facilitates communication among participants providing "a way to talk differently"". (p108)

The reason that Phillips suggests that a more lateral and intuitive complex of ideas is necessary to model a group's thinking processes and support their decisions is because:

"Effective decision making requires a balance between content, structure and process. Content refers to data, information and value judgements that are relevant to the problem at hand; structure shows how items of content are related; and process concerns how content and structure are generated and used in taking decisions. If any of these three elements is neglected, the quality of decision making suffers." (p108)

Taking stock from an applied perspective

As I suggested at the outset of this section, subjective Bayes approaches do encounter some difficulties. Consider what sort of process one would need to deliver the substance of everything discussed above. Consider that these are not seen, on the whole, as alternatives but as "requisite" for effective decision modelling. Notwithstanding the need for a skilled facilitator and a fast computer at all times, there is simply a feeling here that over-engineering is essential to assure success within the normative world-view in the subjective setting.

1.7 What benefits package should decision support provide?

Looking at all the landscapes I have drawn (some rough sketches, some more detailed) we can see the science divided along the fairly classical lines from the laboratory to the field (and sometimes even the field is, critically, just outside a convenient laboratory). There is a lot of nuanced discussion which is really about definition. Whilst we see a lot of applied crossover of the dominant normative basis to the arguably more ethnographic descriptive approaches, there is also a lot of direct challenge in the opposite direction i.e. that descriptive or real world approaches perhaps should define the area themselves.

In the real to life decision theories above we also see a raft of very "methodologically rich" solutions which are problem oriented rather than theory facing. These aim to take decision support into a deliberative, or value oriented, direction and away from logical positivism. Naturally where these use probabilities or combination rules which are statistical, they close the circle again.

As a review focussing on the practicality of applying heuristic reasoning to safety and business critical settings there is perhaps the basis to ask what standards could we apply to a perfect reasoning system. Arguably these could be expressed in a dubious syncretism. The perfect system would go some way to meeting the shortlist overleaf:

Eighteen features of an ideal decision support system

1. Mirrors accurately what people really do but in rationality terms
2. Reflects explicitly the values that are being brought to bear
3. Has credibility with users
4. Is sufficiently comprehensive
5. Uses weighted subjective criteria and objective
6. Allows decision making to take place under conventional circumstances
7. Includes the best available mathematics suited to the problem space
8. Involves some formal logical rules for decision closure (choice, acceptance)
9. Phrases propositions in favour of long term memory use
10. Remains transparent to and explicable by decision makers at all times
11. Rests on semantic and narrative descriptions of problems
12. Takes cognitive limitations into account
13. Bases decisions on a bridge between induction and deduction
14. Does not have too many variables
15. Is not highly demanding of time and resources
16. Can address small scale few stakeholder problems
17. Is able to identify and counter biases in reasoning, or at least inconsistencies
18. Allows individuals to compare reasoning on a common platform with each other

Eighteen features is an accurate summary of the literature reviewed so far, but it is also a very long list. On closer inspection certain commonalities can be viewed and the list can be revised under some more inclusive headings as shown:

Core concept	Eighteen features of an ideal decision support system
Application / workload	Can address small scale few stakeholder problems
	Does not have too many variables
	Is not highly demanding of time and resources
human centred	Phrases propositions in favour of long term memory use
	Rests on semantic and narrative descriptions of problems
	Takes cognitive limitations into account
Rationality	Mirrors accurately what people really do but in rationality terms
	Is sufficiently comprehensive
	Uses weighted subjective criteria and objective
	Includes the best available mathematics suited to the problem space
	Involves some formal logical rules for decision closure (choice, acceptance)
	Bases decisions on a bridge between induction and deduction
	Is able to identify and counter biases in reasoning, or at least inconsistencies
Values / credibility	Has credibility with users
	Allows decision making to take place under conventional circumstances
	Remains transparent to and explicable by decision makers at all times
	Allows individuals to compare reasoning on a common platform with each other
	Reflects explicitly the values that are being brought to bear

To design a system which faces up to such a short-list for a real-time distributed decision support system is all very well. However since this system is aimed squarely at the prioritisation of risks, we have a second key scientific area to consider. This area of risk will prove equally disputed, equally at times on the horizon of a post modern vortex of theory and counter theory, equally given to reductionist and expansionist world-views.

Chapter 2: Literature review of risk theory

2.1 Summary of chapter two

The landscape of risk is one of the most dense areas in recent social psychological history. This chapter will consider the complex literature of 'risk'. Risk will be shown to have always been complex. The meta-themes of the area can be shown to have shifted a great deal over three decades of research and these are considered.

The prevalent theoretical positions on risk will be compared. This covers risk expressed as: science (e.g. the physical hazards); public policy (e.g. the debate in various safety and environmental arena on acceptable / tolerable risk); social construct (e.g. the scrutiny of science and scientists as part of the debate); psychology of the individual (e.g. the various cognitive and psychometric paradigms) and finally post modern (e.g. the idea that one might hold 'risk as science' and 'risk as emotion' in some kind of tension). The various schisms in the definition risk resulting from it being such a complex, socially constructed idea will also be explored.

Social science will be shown to be the dominant frame for risk research and for risk definition. One key debate in the field is discussed in the light of an early nineties controversy between "scientific" risk and so called "perceived risk". This controversy effectively split the discipline and the effects of that split can still be seen cutting across the literature.

The problem of there being little hope of a unified understanding of risk is an important theme to consider. Major representative works ranging across global economics, cultural theory, social science/philosophy, engineering and health science are explored for an explanation.

What is clear is that these fields are distinguished by asserting a different form for risk and a different method for its operation.

The conclusion will be unsurprisingly that risk is a vitally important applied concept however ill defined it may be. This means that, judged within the context of an the applied system for reasoning in the setting like the ones in this study, a definition is still required. The solution given to that problem in this review is, similar to that discussed for decision support, the application of an innovative form of heuristics.

In this chapter we will also touch on literature in the subject matter areas of the two case studies. This material will reveal a small sample of the pressures the organisations face. This awareness will help to contextualise this thesis.

2.1.1 Summary of key points in this chapter

- Conceptual clarity in risk is elusive and at the same time important
- Risk, like normative decision theory, can have a purely model world theory base
- Risk, like descriptive decision theory, being social in nature must accommodate subjectivity within all its conceptions
- That subjectivity still requires rigour of definition to be a sound scientific application, but the definitions have to be "local" to the problem field
- The schism between real and perceived risk can be shown to be a fallacy
- A clear form of risk which can be applied in all settings would never be available from a synthesis of the extant theory

2.2 What is risk?

What is risk? The briefest scan of the works researching this area will show that we might as well ask what is art, how does one recognise quality, or what is the essential nature of time? The more that is written on the subject, and indeed the more it becomes absorbed into our schools, hospitals, offices and factories as "a thing" which is measured by assessment and to which we are answerable under law, the more post-modern a concept I would argue that risk becomes.

Thinking of risk as post modern is not dangerous. Post modernism itself is of course hard to define, it relates to the demise of any absolute truths into relativity. More importantly, for my use of the term, it relates to the fragmentation and division of all academic subjects into a variety of perspectives - with no 'answers' or agreement. It is in this sense that I want to suggest that risk is a post modern concept. It is not that there is no definition of it, it is that there is so much dispute about it. Seeing it as post-modern therefore does not mean that it is unimportant, quite the reverse. It is very important for any system that wants to measure and use risk, particularly if that system promotes human safety, to attain conceptual clarity.

Extant grouping the major themes of risk

Risk is a very large and conceptually complex area as the core concept of risk takes many forms, a sample like those in the list below could go some way to illustrate that:

Statistical and scientific meanings

- **Statistical likelihoods** can make comparisons from historical data, e.g. the numbers of people killed and injured annually, as complex proportions of those who travel where, when and how

- **Exposure** a concept which although personally qualifying risk still does so with respect to statistical distributions e.g. the actual number of times one flies/drives etc.
- **Technical risk of hazard** derived from humans coming into contact with complex technical (or natural) systems, the hazards these present both in normal operation and in their failure modes
- **Observable track record** implies that objects or processes e.g. safe performance, are subject to empirical rules of observation in time
- **Uncertainty** and the degree to which causes and effects are observable in complex, changing, or novel, systems makes establishing formal 'risk rules' very difficult

Human centred meanings

- **Perception** of risk can be highly paradoxical and will vary with context
- **Culture**, the degree to which decision makers, e.g. in industrial safety, can have the luxury of rigorous application of the frequently changing scientifically rational evidence

This thesis will go on to explore the operational use of a combination of risk and risk concepts like those above in two very different industries. The first industry is safety critical, the National Air Traffic Services of the UK. The other is one of the worlds largest fast moving consumer goods companies, Unilever. To span the potential "risk usage" of two such diverse industries a wide range of risk theories will need to be considered. It is advantageous to have some kind of controlling model (however incomplete) within which to examine them. I will consider the major risk theories, as regards the problem space of this thesis, in five distinct clusters:

1. Risk expressed as science
2. Risk expressed as the interface between science and public policy

3. Risk expressed as a social construct

4. Risk expressed as the psychology of the individual

5. Risk expressed as approaching post modernism

The treatment of these clusters are summarised overleaf:

Risk expressed as science	Identification of physical hazards Description of threat Quantification of probabilistic elements Linked to control behaviours	Factual and physical sciences bias hidden assumptions, closed debate
Risk expressed as science and public policy	Rise of the critique of perceptions Challenges to the authority of science Introduction and legitimisation of subjective judgements and reasoning Plural definitions of risk in the same decision spaces Rise of emphasis on communication Rise of tolerability Rejection of the idea of a single mathematical construct Dispute between the physical and the social sciences	Social and political sciences important, transparent assumptions, truth held in tension
Risk expressed as social construct	Geo-political framework application Culturally constructed and seeks more serious cross cultural explanation Value driven control mechanisms Constructed for society not emergent property from it Multiply framed even within a single society Persistent view that risk is about human health Social consciousness not a private process Places science and scientists within the frame of scrutiny	Social and cultural sciences important, multiple permissible assumptions, open questioning
Risk expressed as the psychology of the individual	Risk is about cognition and problem solving Risk is about a the forces which frame perceptions Risk can be measured along existing psychological lines Risk is a model held by the individual Risk has to refer to the self and others Risk is an expected utility artefact	Cognitive and emotive individual and group processes, behavioural decision theory is key reference point, highly diversified and conflicting schools
Risk as post modern construct	Lay persons' perceptions and judgements carry equal weight with experts Framing risk outside of expert models is valid Emotional and scientific responses to risk in the same person at different times are equally valid Risk in toto is challenged as potentially not meaningful	Questioning is key, plural realities can co-exist, the construct has no fixed meaning

2.2.1 Risk expressed as science

In the 1970s the focus for risk was on developing scientific methods for identifying and describing hazards and assessing the statistical probability of adverse outcomes. "Risk analysis" became a scientific discipline which focussed on threats to health such as chemical exposures, road traffic accidents, and disasters. By the early 1980s, risk analysis had evolved to give rise to "risk assessment and risk management". These disciplines started a focus on hazard and risk factor control. Probability became coupled to an assessment of scale in a classic conceptualisation of risk as probability multiplied by effect (Carter, 1995). More and more focus was being placed on managing risk by reducing the scientific measurement uncertainties of risk. This approach was in pursuit of a clearer definition of the form and function of a risk construct which was still very scientific. (World Health Organisation, 2002).

Risk, according to institutions such as the U.K.'s Health and Safety Executive, (HSE, 1999) and academics who advised and supported them, was a scientific question. Risk became 'the realisation' of the concept of industrial health and safety. If one can adequately identify, rank, manage and monitor all threats in the industrial environment, that environment will be "safe". On a societal level this safety would be expressed as stopping factories exploding and poisoning people, stopping aircraft crashing and so on. This would have to be expressed at the individual level as controls on behaviours linked to the preservation of self and others.

For theorists like James Reason (Reason,1997) the problems of the risks endemic in industrial processes have an industrial solution, that of computational fluid dynamics. Risk, like a gas in a chemical factory pipe, has to be contained and it must not escape through "holes". The identification of these holes and the factors which cause them would lead to better and better risk potential mitigation. "Safety" would be the result. "Human error" remains a key (and ever present) "hole" in such a systems approach, but this too can be quantified (one times ten

to the minus seven in any large engineered system) and can be factored therefore into a "safety equation". Risk, for HSE and other safety engineers acting on this kind of advice, is best assessed by physical science using engineering process control metaphors.

2.2.2 Risk expressed as science and public policy

The 1980s saw the rise of greater government involvement in legislating risk controls. The 80's was also a time of conflict over risk. Challenges were being levelled at the authority of quantitative science to own, describe and measure it. The rebuttal was that scientific predictions were still seen to be rational, while public perceptions were believed to be largely subjective (the so called 'real and perceived' risk debate). Policies to "correct" and "educate" the public however met successful resistance in the era of rising distrust in governments and industry. Scientists too, now in the same spotlight, revealed the high levels of scientific uncertainty that were inherent in the risk field. Challenges insisting on alternative assessments and alternative interpretations of 'risk' gathered momentum in environmental and public health debates.

Commenting on these debates the WHO concluded that governments and politicians had a major role to play in handling conflicts over trust in risk policies by promoting open and transparent dialogue.

"The so-called scientific or quantitative approach to health risk assessment aims to produce the best possible numerical estimates of the chance or probability of adverse health outcomes for use in policy-making. Although high credibility is usually given to this approach, how valid is this assumption? Why is this approach often seen as more valid than the judgements made by the public or social scientists?" (WHO 2002, p30)

The WHO comments how *"in practice there are considerable difficulties in making "objective" decisions at each step in the (risk) calculations."* (My brackets). Thus any risk modeller has to adopt (and declare?) a specific definition of risk and needs to introduce into the model a series of more subjective judgements and assumptions (Carter, 1995, Slovic, 2000).

By the early 1990s, particularly through the failure of scientific approaches to control and explain risks, it became accepted that risks had to be understood within the larger social, cultural and economic context (Gifford, 1986, Pidgeon, 1992, Stern and Fineberg, 1996). Risk became seen as embedded within societies and their cultures. Determining that individuals perceive and control risks within a cultural frame was not actually an entirely new idea (Douglas M, Wildavsky, A 1982).

Increasing disillusionment that the "lifestyles" approach to risk, based on health belief models (see Janz and Becker, 1984), had singularly failed to yield sufficient behavioural change (in the global context of the rapid emergence of / failure to stem the rise of HIV/AIDS) further weakened accepted scientific norms. This became an era where the public and NGOs, particularly those in the environmental movements, became better organized and more effective lobbyists.

In post BSE Britain, improving public risk communication, rather than the science of its measurement, was seen as the essential focus of the debate. Powel and Leiss in their book 'Mad Cows and Mothers milk' commented:

"Good risk communication practice exists in the zone that separates the languages of expert risk assessment and public risk perception. . .both languages are necessary, because the daily business about managing risks – both the personal business of individuals and the social

allocations of risk reduction resources – cannot be conducted in either one alone'. (Powel and Leiss, 1997, p29)

It was in this intellectual environment that, in 1992, the Royal Society had released a report entitled 'Risk Analysis Perception and Management'. This was a further investigation into risk following on from their 1983 report entitled 'Risk Assessment'. A key difference obvious from the titles was in the proportion of social scientists now present (and recognised) in the risk area. The 1992 report's intent of providing a state of the art review of risk was somewhat doomed by one argument, how does one define risk, as excerpts from an early chapter on risk definition show:

- *"The word risky is undefined, and it not to be used as a synonym for dangerous. All risks are conditional, although often the conditions are implied by context rather than explicitly stated.*
- *Although detriment may represent the only numerical way of comparing different events associated with the same hazard, or the combined events of different hazards, the fact that any such comparison is an arbitrarily weighted total of incommensurables must never be forgotten.* (p3)
- *There are serious difficulties in attempting to review risk as a one-dimensional objective concept. In particular risk perception cannot be reduced to a single subjective correlate of a particular mathematical aspect of risk, such as the product of the probabilities and consequences of any one event."* (p7)
- *Given the essentially conditional nature of all risk assessment, one should accept that assessments of risk are derived from social and institutional assumptions and processes; that is, risk is socially constructed"* (p7)

These and other factors led to a philosophical shift from a focus on absolute risk to that of "tolerable risk". This debate is credited as having began in earnest (in the U.K.) when Sir Frank Layfield's 1983-85 public enquiry into Sizewell B (O'Rierden, Kemp, Purdue, 1988) concluded that he disliked the term "acceptable" because some risks, given their benefits, were "tolerable". One could foresee a *"willingness to live with a risk to secure certain benefits"*. Thus the ALARP ('As Low As is Reasonably Practicable') principle was born.

The ALARP principle was being taken forward as a matter of public policy by the U.K. Health and Safety Executive. They stated in 1991: *"the judgement on what is tolerable is not a scientific but a political matter"* (HSE, 1991, p13). This reasoning would give rise to a competitor notion, "the precautionary principle" (that one should not go ahead with a risky venture until reasonable proof of its safety was scientifically available). Pressure groups would use this principle to place the onus back onto the risk scientists to provide proof of safety as a necessary condition for a course of action.

What is important to see is that the debate between science and public policy made risk into a more complicated interface between the two. The science was now firmly in the social science arena (not solely the physical or statistical) and the public policy debate opened the door to a very wide set of publics to discuss the purpose of risk reasoning.

Fischoff et al had originally suggested there may be three generic institutional approaches to 'acceptable risk' (Fishchoff et al, 1981). These were: professional judgement (by individual or using 'standards'); formal analysis (e.g. cost benefit, decision analysis) and 'bootstrapping' (revealed preferences from hazard tables etc.). A decade later, the authors in the Royal Society's report suggested there were now seven criteria for acceptable risk, these were:

- Comprehensive
- Logically consistent

- Practical
- Open to evaluation
- Politically acceptable
- Compatible with existing institutions
- Conducive to society learning about risk
- Improving of decision making in the long run

The public policy debate had created the need for an expanding social science framing of risk which was nearing absurd levels of complexity.

Risk science had suggested a controlled application of a well defined measurement concept in well understood decision problems. Risk social science, particularly at the public policy level, suggests the need for any measure to take account of tensions between values (agreed and disputed), agreed facts and disputed facts. Risk science had estimated risk as a statistical value. Risk social science somehow had to be able to reflect these values in a more complex 'consistency'. Importantly that consistency was a construct and not a number. *"There are serious difficulties in defining a single measure of objective risk itself "* (HSE, 1991, p90)

The problem, according to Adams, with the well intentioned Royal Society's approach (Adams, 1995) was that between 1983 when it published the *"authoritative, confident and purposeful"* Risk Assessment and 1992 when this second report came out, something tectonic had happened in the world of risk, or more precisely in the scientific perception of risk.

Adams relates the 1992 report thus:

"Although it was published by the Royal Society, the Society was sufficiently embarrassed by its contents to insist in the preface that it was "not a report of the Society", that "the views expressed are those of the authors alone" and it was "merely a contribution to the ongoing debate"".(p7)

Adams suggests that by 1992 the Royal Society was no longer capable of taking a collective scientific view about risk. He attributes this squarely within an irresolvable dispute between the social and the physical sciences. To his view they simply could not agree about the meaning of risk because risk could not be a unitary concept.

2.2.3 Risk expressed as a social construct

The debate on the social construction of risk spans the entire time-period of the changes in public policy outlined above. Since the early eighties, serious effort had gone in to looking at the construction of this concept and establishing the arena of its impact. This has a range extending from geo-political to socio cultural dimensions. Some of the key theories are discussed.

Geo-political Risk: Risk, according to the sociologist Ulrich Beck, (Beck, 1986) is a global force, like economics. To him it is a "rising" force which will become the dominant factor in shaping the future of the human race. He argues this bold point extremely effectively when he discusses the way in which science and technology has harnessed forces which can incinerate or poison the whole human race. The core essentials of life, he argues, like food and water, are no longer subject, as they were historically, to economic forces which dictate who has control over access to these. Rather, the global forces of risk will dictate whether all food and all water (and all air for that matter) are safe for consumption by all human beings irrespective of geographical position or economic wealth. Risk, in Beck's overall conclusion, will be satisfied in what he calls "reflexive modernity": that is where humanity better harnesses science and technology to learn all the faster how not to threaten the geosphere. For Beck risk is assessed by awareness of social phenomenology and dominated by science and technology.

Culturally Constructed Risk: Risk, according to Mary Douglas and Aaron Wildavsky (Douglas and Wildavsky, 1983) is essentially a cultural phenomena. It is value driven and is used primarily for control of a stabilised society which is normative. What any society believes about economic forces and indeed the "myths" it holds about the functioning of the natural world, are a a control mechanism of the powerful. How things go together, and what to fear therefore, is defined by these agents according to cultural theory.

The current debates in 2007 about the realities of global warming and climate change would fit very nicely into such an interpretation. For Douglas and Wildavsky one must understand who is using risk and what they are using it for to understand what it is. Risk can, in their view, only be seen through the filter of an anthro-centric world which may or may not have any contact with the real world. For Douglas and Wildavsky risk is assessed by those who have social power and is dominated by myths of nature when these are dressed as indefatigable scientific truth.

Social Scientific Risk: This debate has been touched on in the previous section. Risk, according to Statistician and Geographer John Adams, (Adams, 1995) is a social construction. Borrowing what he considers the best from cultural theories he uses the statistical observation of society to prove that risk can only be seen as socially constructed and multiply framed within that. This construction and framing remains elusive as it is somewhat dynamic.

Like the cultural theorists Adams suggests that risk is about societal control, for example over our beliefs and behaviours in driving on the roads. Adams explores what he sees as society' odyssey with risk. This is taken from the point of view of who it is that wants to control the definition of risk. In particular he reviews the desire to make risk in the image of other measurable values with which society is more familiar and comfortable such as assessing the

financial value of "a risk". Risk, in Adams view, has to be assessed through the same filters of any other social construction and will be dominated (in the west at least) by utility and measurability.

Societal Welfare Risk: Risk, according to the World Health Organisation (WHO, 2002) risk is a problem for society. Getting a better understanding of how risk works politically, economically and socially and where the key risks are in the food chain and in the health promotion cycle of modernising society is seen as the key task. The WHO is not philosophically very far from Beck in this regard. To the WHO, risk has to be dealing primarily with accepted and obvious risk to life, be assessed through measurable human health realities and be dominated by governance decisions.

Socio Cultural models: A back-lash against the dominance of cognition and decision theories these models consider intuitive and emotional aspects of risk. Cognitive type models would suggest that risk is driven by private understanding, socio cultural models would emphasise the product of continuous process of interaction in a social context. Douglas (Douglas and Wildavsky, 1982) developed the use of a world-view approach and see different social groups as selecting different risks to pay attention to. This would explain differences in perception and those which they are pre-disposed to select.

The agent as selector of risk (paying selective attention) is influenced by cultural bias i.e. worldview and the arrangement of social relations across five axes: Hierarchical, egalitarian, individualist, fatalist and hermit. So people do not focus on risks primarily to protect their safety but as a means of protecting their way of life. Douglas has suggested that the heuristics of risk can be re-phrased not just to be individual cognitive strategies, but to be a medium to

clarify options and expectations related to other individuals and thus predict what they will do along agreed cultural values.

Many failings of such a soft approach have been pointed out, for example it doesn't explain how culture changes over time and it perhaps oversimplifies things. Its key strength has probably been to drive a pluralist wedge into the hinges of the scientific doors. It has perhaps helped theorists like Wynne (Wynne, 1992) to subject science and scientists to the same scrutiny and critique normally reserved for lay people in scientific models. Wynne suggest that scientists too are biased, selective, normative etc.:

"What is taken by technical experts and policy makers to be an irrational rejection of scientific information may instead by a rejection [by the lay populous] of naïve assumptions [within expert models] about an ideal world, both social and material, which is embedded in the 'expert' model of risk taking". (Wynne 1992;p.32)

Cross Cultural Models: Boholm attempted to make an overarching review of the cross cultural aspects of risk perception over the twenty years up to 1998 (Boholm, 1998). As a result he encouraged deeper and more moral and inclusive forms of research. Boholm felt this research should face up to the richness of cross cultural issues rather than trying to sublimate them in trivialising generalisations. He was frustrated by the psychometric paradigm and its over emphasis on measurement. He was also frustrated by the fact that extant cross cultural studies:

"form a heterogeneous field of research, with little agreement on basic theoretical issues, an low consensus on problems in their research or on methodology. . .as a matter of urgency, we should be systematising results and theoretical frameworks, thus generating new research which will lead to more powerful statements about the perceptions of risks as social and psychological phenomena." (p153)

Boholm was heavily critical of circular social science argumentation where the answers to questionnaires are corroborated against "superficial characteristics of the society in question". Linking these characteristics to objectively observable phenomena pointed, as far as he is concerned, to merely asserting that social science has found exactly what it might have expected. Boholm suggests a new paradigm is needed:

"about the way in which risks are embedded in the social fabric, taking into account 'conceptions of morality, equity, justice, and honour; religious doctrine; ideas concerning sovereignty; property, and rights and duties; and aesthetic values and what constitutes quality in life' (Rappaprot, 1996: 65)."

With this he wishes to do away with the:

"fallacy of unrestrained empiricism, producing new empirical results accompanied by trivial explanations for the sake of it (Fauchaeaux, 1976)" (p160)

2.2.4 Risk expressed the psychology of the individual

Risk, irrespective of the prejudices that have been filtered into our scientific world-view and assumptions base, is a human construct. Risk therefore, whether it is through science or through experience, is always perceived. A number of critical perception theories have been put forward in the psychology of individual risk. In 1999, the UK Health and Safety Executive, conducted an excellent and wide-ranging review of these entitled: 'Risk Perception and Risk Communication: a review of the literature'.

That review suggested that to understand risk and how it functions one must question how the construct is appropriated by individual human agents. This appropriation however can be seen through a number of theoretical filters. An in-exhaustive list of these would include:

Cognitive; Psychometric; Mental Models, Value Expectancy Models and their application in industrial "safety culture" models.

Cognitive: The cognitive paradigm accepts that people actively evaluate risks in some from of costs benefits trade off. Until more recent research (see earlier critique on elicitation and heuristics) this approach emphasised lapses from optimal rationality. Tversky and Kahneman's laboratory studies (Tversky and Kahneman, 1974, 1981) were highly influential showing that human decision making did not follow Bayesian logic and that much of human decision making is typified by biases, errors and misconceptions. Tversky and Kahneman and others suggested "cognitive heuristics" were being used by decision makers. These heuristics created a useful reduction in the cognitive processing load for a decision. However, the heuristics were prone to 'misconception' and 'misapplication' which lowered the decision quality.

A number of other theorists (Slovic, 1987., Lindberg and Frost, 1992., Doyle, 1997.) took the area forward to conclude:

- sources of bias can be directly introduced from the way risk information is portrayed
- framing can be a key factor e.g. where there is a generalised preference for options phrased as gains rather than losses.
- ownership of the risk outcome is an important mediator
- people tend to see risks as discrete exposure rather than increased exposure.

These cognitive framing issues had achieved the status of normal science of the 90's but remain in dispute.

Psychometric: Perceived dread and perceived familiarity are key factors in expressed preference for risks. Risk in this approach is considered inherently subjective. Risks are only

seen through the social filters of the way people apprehend them. For example, people underestimate likely risks and overestimate unlikely (Lichtenstein 1989) and people are better risk judges when rank orders were used (e.g. Pidgeon 1992).

The underlying rules which people might have be using to understand risk were written from a factor analytic approach leading to sets of widely used 'key factors' of personal risk assessment as being:

- Perceived dread (how much is the object feared)
- Unknown risks (to what degree is the object alien to comparison)
- Number of people exposed (to what degree does this potentially impact)

Cross cultural differences have emerged in these factors however (Englander, 1986) and there was much criticism of the pre-selection of hazards by the researchers ensued (in much the same vein as the pre-selection of decision problems in the related heuristics research). Slovic (Slovic, 1983) attempted an organic approach emphasising the inherent weaknesses of psychology experiments (sample size, representativeness, control of variables etc.) and the factor analytic approach which was so frequently used. The unsurprising conclusion one comes to when the core methodology is undermined in this way is of course:

"...the concept of risk means different thing to different people" (Slovic 1986)

Mental models approach: This approach explores people's declarative knowledge through free elicitation techniques and compares experts and non experts through the development of mental maps. The study of errors and biases in this context has shown that the inferential processes of lay people are different to those of experts. Where experts will differentiate between correlation (a relationship) and contiguity (the juxtaposition of unrelated events) lay people generally did not. Lay people tended to class all cases more deterministically this is perhaps because people want to understand and control threatening uncertainties. The

tendency, apparent in this model, to reference lay views to experts remains very strong as does the assumption that the expert has the correct view.

The correctness assumption has been challenged by Wynne (Wynne et al, 1992). Wynne tried to expose misconceptions and biases also present in the expert mental models. Disunity in science over things like food scares was seen as contributing to a new view of science, a view of distrust and an erosion of the perceived credibility of the sources of risk communication (MacGregor and Flemming, 1996)

"... lay models of Toxicology, and of chemical risk, are not only less complex [than the experts] they are primarily composed of beliefs, attitudes; perceptions and impressions that are loosely organised according to intuitive principles" (MacGregor and Flemming, 1996)

Value expectancy models: These models were aimed at unpacking cautionary motivation and self protective behaviour. Their application is to do with health promoting behaviour with a central notion of perceived vulnerability. This work relies on a strong (but contested) relationship between behaviour and attitudes and is underpinned by behavioural decision theory. In value expectancy models risk behaviours reflect a conscious decision process involving the utility of outcomes. The individual wishes to minimise harm, through action based on perceived seriousness. That motivation is said to be based on perceived likelihood of event. Benefits must be weighed up in terms of costs. There are three dominant theories in this area:

- Theory of reasoned action; Fihsbein and Ajzen, 1975
- Theory of planned behaviour; Ajzen, 1991
- The health belief model; Janz and Becker, 1984

There is considerable overlap in the models as all three theories have a common ancestry, being based on Subjective Expected Utility Theory. The theories also share a belief that subjective cost benefit assessment takes place in the form of some sort of function.

Later research (Harrison et al, 1992., Van der Pligt, 1998.) heavily criticises these approaches. Problems highlighted included "cognitive framing issues" relating to framing, misinterpretation of probabilistic information and cognitive availability issues e.g. over-estimation of sensational or dramatic risks (Van der Velde et al, 1994). Likewise, perceived susceptibility was mediated by the ease of visualising yourself as a victim. The 'self and other referent' bias and 'unrealistic optimism' are both cited as complications. These are both complex psychological entities with attribution biases which preserve the self image. Asking people to do comparative risk assessments on themselves seems to trigger social comparison processes rather than risk or health precautionary assessments (Van der Pligt, 1998).

Safety Culture models: Safety culture models are more of a theoretical emphasis than a new theory. These models emphasise the importance of culture in the workplace when trying to identify what makes it safe. The role of culture and sub cultures (e.g. professional groups) has to qualify how risk functions. This tends to be a framework type approach. There is support for the idea that people actually have an accurate idea of risks in industrial contexts (e.g. Fleming et al, 1998) although this perhaps relates best to more physical risks found in high hazard environments. That accuracy may drop in more complex environments.

How risk is assessed and by whom is the key issue in a safety culture assessment. Managers might be pressed by alternative worries which places them farther away from the centre of the risk itself. Thus their decisions may underestimate risk. Safety culture research, being

predominantly psychometric in nature, suffers the same criticisms of those approaches already discussed (attitudinal measurement and drawback of factor analytic approach).

The HSE (HSE, 1999) comments that it is not the conceptualisation of risk which is at issue amongst differing social theorists, this is generally accepted within the psychometric tradition (Fishchoff, 1995). It is the status which is ascribed to lay knowledge and the extent to which subjectivity is accepted as being present which differentiates the 'psychology of the individual' theories of risk.

What is interesting about almost all of these psychological theories is how heavily they draw from the concepts of wider decision theory. Some concepts are taken directly from decision and game theories. Notice a common emphasis on a range of human cognitive biases, for example the failure to understand set theory. Subjective attitudinal approaches emphasise cognitive emotional framing (e.g. the impact of dread) as if it were a factor in a risk judgement calculation. Inferential process approaches compare expert and non expert judgement on (largely probabilistically framed) risk choices. Health beliefs and value expectancy approaches emphasise conscious utility-driven behaviours in risk choices. Many of these approaches point out lay people are inconsistent in using statistical and scientific judgements to inform behaviour.

2.2.5 Risk expressed as approaching post modernism

Post modernism as I have already pointed out is a complex concept concerning the collapse of academic study of an area into a arena where there are no answers, only competing conceptualisations. As has been seen risk can be argued to be suffering from this problem. A further common use of post modernism asserts the validity of an individual's "own" reading of reality. Risk in some more recent research has also adopted this flavour.

Lay Risk: Theorists like Andy Stirling and Brian Wynne have made a lot of headway with the notion of 'lay persons risk'. They have argued that this risk is equally admissible in applied industrial problem solving e.g. in the location of waste incineration facilities, the response to genetic modification of foodstuffs. This is arguing for a more post-modern understanding of risk. Stirling and Wynne and theorists like them are saying risk cannot be conceptualised outside of the individual's experience and that groups of individuals should be allowed to set the essential frame for their risks outside of "expert scientific models".

For these sorts of theorists lay risk has equal weight with scientific risk. Risk is assessed as a group rationality process (with scientists as one player) and dominated by a relative expression of preferences surrounding commonly agreed fears. The problem for governments and industry in particular is that this is a break with the authority of science to determine the factual truth of a situation.

Emotional Risk: A relatively recent innovation in the risk area which is appropriate to add to this section on post modern readings is Slovic's notion of "emotional risk". In this attempt is to tease out risk behaviours in a more classically philosophical taxonomy (Slovic et al, 2002) he would argue that we are none of us, scientists or lay people, bound by any one appropriation mechanism for risk. Rather, we sometimes see risk as science and behave appropriately. We sometimes see risk as emotion and then we behave differently. In the latter case we are not behaving inappropriately to contest or reject scientific views on environment, safety etc., it is just that our "risk affect" is the focus of our rationality for reasons which are entirely explicable.

Risk Communication: Once the model for understanding risk has been selected another model for communicating it will be important. As we noted earlier the late 1990s saw a rise in the emphasis of risks being adequately communicated suggesting that this was not a trivial matter of disclosure to a receptive audience. Two debates around risk communication which unsurprisingly mirror the debate around risk these are:

1. The idea that risk communication has no meaning
2. The idea that risk is the key currency to educate the health perceiving public

I've chosen two more idiosyncratic sources from among the many alternatives to examine this debate. My reason for doing this is two fold. First, these considerations have the same essential qualities to offer the debate as any of the more mainstream ideas, as they are well reasoned and adequately contextualised. Second, using them as sources highlight just what degree of penetration the risk debate has had:

In two papers, both published in 1999, Jack Dowie was an interesting dissenting voice on the need for risk at all. His paper "Against Risk" he says this:

"'Risk', whether used separately or in conjunction with other terms (as in expressions such risk assessment, risk factors, acceptable risk, and risk communication) is an obstacle to improved decision and policy making. Its multiple and ambiguous usages persistently jeopardize the separation of the tasks of identifying and evaluating relevant evidence on the one hand, and eliciting and processing necessary value judgements on the other". (p57)

Again in 1999, this time particularly focussing on risk communication being meaningless, he says that risk is a "conceptual pollutant":

". . .it encourages people to assume that they know what they are talking about when they use it – but, much worse and much more significant, to assume that they know what others are talking about when they hear or see it used. It is a highly dangerous chimera which can be dispatched without loss".(p42)

Dowie, in no uncertain terms, feels that risk used in a decision analytic setting should be

"replaced by a term, or terms capable of carrying the requisite analytical burden"

These pleas were published in journals whose names included risk. They were published at a time when risk was being more and more highly diversified and nuanced. History found in favour of this diversification and our desire to carry on using this term coupled to "communication" remains. The problems of defining the communicative intent persist also.

Risk has clearly become a plurality which conveys its diverse meanings to a range of publics who have internalised some part of it. Scott Ratzan, in his editorial for the Journal of Health Communication in 2003 entitled "Making sense of Risk" took a very pro-science and anti-precautionary principle approach. He tours medical advances which would not have been possible if the precautionary principle (more and more he risk regulatory norm) had held sway. He suggests that a precautionary approach to risk hampers progress on stem cell and GMO research and a host of other areas and is "threatening the future of scientific innovation". After highlighting an embattled government (over SARS or Acrylamide) and the "media frenzy" which leads to public mis-trust his plea to health communicators is this:

"Al all of us are health communicators, we ought to be aware of the role we play in the proactive and reactive approaches to risk. Such areas of Knowledge Management, Knowledge Sharing, and Knowledge Utilization are amongst our arsenal to approach a wary public." (p399)

Recognising the need for "discrimination between evidence and values" he goes on in his short address to cover risk communication (on health) which must:

- Tailor language to needs and interests of users

- Be clear reliable relevant and credible
- Sensitive to cross cultural variation in risk acceptability
- Sensitive to individual variation in risk acceptability
- Discriminating of voluntary and involuntary risk
- Sensitive to differing emotional responses

Hopefully, the strain of ever producing adequate risk communication which just this one sub-sector list requires is a point in itself. Ratzan has provided yet further evidence of a post modern, individualistic reading of risk (and risk benefit analysis).

2.3 Risk and Decisions for whom?

The National Air Traffic Services of the UK and Unilever are "corporations" and need to have a stable, useful and corporate concept of risk to reason from and make policy within. Whilst the actual work of these organisations is not the subject of this thesis, it is appropriate to very briefly and incompletely position some headline themes:

2.3.1 Risk and the National Air Traffic Services

NATS is a provider of air traffic control services for the mainstream of air traffic in the UK. Much of the literature surrounding this sort of operation is, not surprisingly, fascinated by the preservation of safety and the reduction of (classical) risk in that environment. This is done through safety technology. It has been argued (Arthur and Sonander, 2000) for NATS that there has been a tendency to design this safety in isolation:

". . .safety is in a complex and dynamic relationship with the whole system utility, within which it has plural definitions and rationality."(p3)

NATS, like its counterparts around the world, was somewhat worried by trends in air traffic such as exponential rises in air travel, the speed of new technology and growing competitive and commercial demands on historically single-minded (safety) air traffic organisation. A key analysis of the situation in America by T.S. Perry had sent some early shockwaves (Perry, 1997). He predicted:

"If the U.S. air transportation system does not change in any significant way, there could be a major aviation accident every seven to 10 days." (p19)

This, Perry was careful to unpack, was simply a statistical prediction in the face of burgeoning air travel and a statistically flat accident rate for over 15 years. He also saw that the historically safety critical nature of air traffic system had created unintended side effects of technology which couldn't change for fear of knock on effects.

The "host" system for U.S. air traffic control, at that time, contained half a million lines of a software language (jovial) which was no longer in use, ergo there were no experts. This computer, at the centre of detecting all safety critical aircraft data to prevent collisions, had only 16MB of RAM, and this was in 1997. Quoting from an air traffic controller interviewee Perry points out that this software was now so old and "patched" that it had to be left alone:

"There are so many patches, no-one knows how it works. We can't change anything, no-one dares touch it. . ."

The approach to this problem was to build better and more powerful peripheral systems to support the ageing central one. Perry points out what a short term risk strategy that could be.

Perry goes wider in his substantial paper to develop a detailed analysis of the possible technological and ethos changes facing air traffic in the coming decades e.g. the proposed introduction of free (not controlled by air traffic) flight, the changes to air traffic technology and so on. What his particular paper did at this time was to 'raise the stone' on the impeccable idea that safety was in safe hands and free from commercial consideration:

"Pilots throughout the world think free flight is a great idea, if implemented correctly. Airline managers are enamored of the concept because it has the potential of saving billions." (p33)

In July of 2000, the European Organisation for the Safety of Air Navigation published an investigation: "Investigating the Air Traffic Complexity: Potential impacts on workload and costs". In this project they attempted to develop an algorithm to compare air traffic centres across Europe equitably on the effects of their traffic, airspace and controller workload complexity. Notice two things, the assumption that this was not only possible but that an scientific algorithmic approach was the desirable solution space. The report concluded that complexity was now at the stage where it too must be measured:

"complexity has to be incorporated if meaningful and cost effective metrics are to be developed." (p27)

Colin Chisolm, the then Chief Executive of NATS, in his 2001 strategy suggested that the pressures from sustained traffic growth, the transition to a brand new air traffic centre (Swanwick), the transformation of NATS into a Public Private Partnership will:

". . .all place demands on the company. The plan, therefore, is centred on a programme of improvement and development, supported by mature and effective safety management processes and underpinned by an increasingly comprehensive analysis of the sources of risk."

Notice the idea of an "increasingly comprehensive" response to risk. It was that idea that led to my research into less conventional human factors in risk and decision making being commissioned.

2.3.2 Risk and Unilever

Unilever is one of the world's largest Fast Moving Consumer Goods companies, or FMCG for short. It has two large product categories in foods e.g. Lipton tea, Walls ice-cream, Flora margarine, and in home and personal care e.g. Domestos, Dove and Lynx-axe. It owns and manufactures some of the largest and most popular brands in the world operating manufacturing and distribution in over 100 countries. It is a "corporate giant".

Reputation management

So, other than stock market and supply chain considerations, why would Unilever have an interest in risk? Bakan's book "The Corporation" (Bakan, 2005) is an example of the "paper-back army" who routinely attack large companies and engender distrust in them. Tracking the rise of corporate social responsibility, which Bakan considers a smoke-screen to

hoodwink the public into supporting the psychopathic tendency of companies to pursue power and profit, he says this:

"Despite this shift, the corporation itself has not changed. It remains. . .a "legally" designated person designed to valorize self interest and invalidate moral concern. Most people would find its "personality" abhorrent, even psychopathic in a human being yet curiously we accept it in society's most powerful institution." (p30)

In its report "Power Hungry: Six Reasons to regulate global food corporations", Action Aid (a Non Governmental Organisation) suggests, whilst citing direct examples of Unilever alleged unethical practices:

"Global food companies have grown too powerful and are undermining the fight against poverty in developing countries. They are draining wealth from rural communities, marginalising small-scale farming, and infringing people's rights."

Conversely in the World Wildlife Fund website one finds:

"Unilever, with advice from WWF, is taking a global lead in the development of sustainable agriculture."

Unilever works in close partnership with WWF and is in regular factual dispute with Action Aid over its allegations and its tactics. However, in its report "Still Dirty: A review of action against toxic products in Europe" The World Wildlife Fund still holds a position that :

"Chemicals are an integral part of modern life: everyday products from computers to food packaging and detergents contain chemicals which can contaminate our environment and our bodies. Few have been adequately assessed for safety, yet exposure to these substances may be a key factor in rising rates of cancer, declining fertility and other reproductive problems affecting western societies".

Food safety

In his review of trends in food safety Korthals suggests that there are fundamental problems for global companies like Unilever (Korthals, 2004) because:

"The definition of food safety and risk differs by culture and age group. A certain product can thus be identified as hazardous and unsafe within one culture, but not within another. This cultural multiformity of scientific and lay perceptions of risks and food safety leads to problems when everyone can appeal to the right to a personal definition of risk" (p503)

In "Risk Management in Post Trust Societies" (Lofstedt, 2006) the wider changing face of trust in industry, governments, regulators and other organisations in the public apprehension of risk is explored. It concludes *" an increasing decline of public trust in both local and national policy-makers"* (p108). With this decline he also notes trust is *"increasing towards other groups most notably environmental NGOs"* (p109)

Throughout his book Lofstedt explores the relativism with which scientist, technocrats, civil servants and national and supra-national regulatory bodies are viewed in the matter of risk and the regulation of large companies. Who to trust and whom to believe over matters of national and global importance in responding to risk, modernity and globalised industry, Lofstedt and others will argue, is a very variable and unpredictable space. To be a key player in that space companies like Unilever must understand and skilfully manage a new portfolio of corporate reputation risks.

It was in this growing realisation that Unilever set up its global Unilever Issues Steering Group. This group commissioned me to build an issues prioritisation process and tool to track and support the management of global issues.

2.4 Conclusion from an applied perspective: Which risk model to prefer?

What is highly evident here is that risk is a pervasive but enormously diverse concept, all of the mainstream models for its action are criticised in one way or another for omitting elements of each others framing and omitting (non western) social and cultural differences in their schema. Risk offers a banquet of overlapping sets of motivators for human beings who must measure it conscious of discontinuity between perceiving, managing, communicating or reducing the inherent risk. This measurement takes place under the need for any given measurement proposition to make sense also.

Risk has to be assessed by examining and by contextualising belief and is dominated by the conflicting stories that goal-oriented persons and groups of persons tell themselves about it. These individuals tell such stories in the face of other individuals who might appropriate the same nomenclature to challenge or undermine the authority or acceptability of the primary goals. If one is going to make use of risk concepts in an applied field, a question looms large: Which model to choose?

At the beginning of this section I suggested that risk is now a post-modern concept. In my usage I wanted to say that risk suffers from academic fragmentation and division into a multiplicity of perspectives - with no 'answers' or agreement. As the evidence here shows I can quite firmly back such an assertion, doubly so if I want to take risk and start applying it (as others have before me) to real situations.

Where that situation, like one of the cases in this thesis, is actually about the loss of life of hundreds of people, and where the risk in question is a new formulation to compliment the contribution of established "risks" already at work, it is very important to attain conceptual clarity. What is abundantly clear is that this clarity is not available from a synthesis of the

extant theory. If we were to adopt therefore the (questionable) syncretism which we used in the case of decision theory earlier, perhaps we could attempt to synthesise at least some governing principles into a similar qualitative framework for risk. An idealised form of risk, arguably from the debates above, would have nine attributes:

1. Risk is about measurement and management
2. Risk has to convey multiple perceptions
3. Risk has to be about control features
4. Risk ownership and risk framing are key essentials
5. Risk is about communication
6. Risk should be able to inform actions and influences
7. Risk should focus on culturally important things
8. Risk should contain a value idea and have an affect component
9. Risk is about decision making

These attributes might further cluster under two main headings

Core concept	**Nine features of an ideal risk system**
Application	Risk is about measurement and management
	Risk is about communication
	Risk should be able to inform actions and influences
	Risk is about decision making
	Risk has to be about control features
Perception	Risk has to convey multiple perceptions
	Risk ownership and risk framing are key essentials
	Risk should focus on culturally important things
	Risk should contain a value idea and have an affect component

2.4.1 Concluding remarks on the literature

What is obvious from this whole review is that risk, heuristics, decision making and human reasoning have a number of common platforms which are typified by debate, not consensus. Not only are the debate platforms similar but many of the researchers in one area appear in the other. Point number nine above remains key. Risk has its expression, according to the prevailing literature, not so much in hazard quantification or the summing of fears any longer (although there is a clear case for this in industrial process control), but in discriminating which (rationally coherent) decisions people do or don't take in consequence of quantified hazards etc. Not surprisingly, although it is not phrased in these terms, elements of

normative, descriptive and prescriptive theory for risk are clear in the literature. Also not surprisingly risk theory is demonstrable in its disunity to the point of post modernism. At the one extreme statisticians are offering to calculate the risk of anything at all with a universal formula, at the other philosophers are still arguing over mind body dualism.

Towards an applied heuristic decision system for risk

In the debates above I have attempted to find a grounding set of ideas for a risk decision support system in mainstream decision theory, in mainstream risk and in mainstream heuristics.

The theoretical position for the use of heuristics contains such heavy caveats of definition that it is simply too dense an idea to use from a utilitarian perspective without significant revision.

The theoretical position for the use of risk contains the necessity of choice from a multiply overlapping and, as yet, unresolved, concept over which no definitive position can defensibly be assumed.

The theoretical position for the use of decision theory is certainly a challenge, not least because much of the theory which might apply to a reasoning system, even that which claims to be applied, is still based in a paradigm of definitional and intellectual control over "a decision" rather than "reasoning" per se. The closest candidate remains a revision of subjective Bayes attribute elicitation.

That said, at the general level I think I have perhaps achieved four things. First, I have demonstrated that this is an impossibly complex and contradictory area of science to sell to an applied audience. Second, there is a sort of research-based Gestalt by which one might

qualitatively guide and judge a decision support tool. Third, there is second, Gestalt for considering qualitatively whether its risk elements are sufficiently meaningful. Fourth, this area is ripe for an applied innovation aimed not so much at simplification as at "elegance".

The context for the innovation has to be that there is a real world where dealing with risks and their impacts is not the core business of most industry. The experts in industry who have to deal with risks are often not risk experts but domain experts who have to master a contextualised form of risk. Professional decision making involving risks will tend to form around pre-existing business decision making processes and argument forms. The persistence of beliefs in hunches, luck and judgement in the execution of expertise has to remain in any attempt to frame decision support.

The innovation itself should be able to describe a system to support decision makers (who are not the subject of theoretically aligned research) in making better and (to their own standards) more rational decisions about their risks (formulated in their own reference grammar for that). I think it remains possible to develop a hybrid idea of decision support and risk utilising heuristics to take this field in an innovative applied direction.

In the light of the highly fragmented and unresolved nature of the risk and decision sciences, it would not be wise, or indeed possible, to build, in any global sense, an expert risk system for the National Air Traffic Services, and thereafter for Unilever. I decided in both cases to take a chance on entrusting the expertise to the subject matter owners. This left me free to build an upstream system focussing solely on improved reasoning, enhanced rationality of decisions and transparent communication of reasoning rationale. This would form around an exposition of the existing natural heuristics of these two decision maker communities.

This would not therefore be an expert system, it would be a system 'for experts'.

Chapter 3

3.1 Summary of chapter 3

Two pieces of applied research were undertaken with the National Air Traffic Services (hereafter NATS) supervised by a steering committee of NATS experts. These case studies investigated risks potentially introduced to the UK air traffic system through human decision making. The decisions primarily surrounded new technology projects and strategic planning.

The first project (Hurisk 1) was a highly explorative piece of work covering nine small case studies into a representative range of NATS decision spaces and project types. A hypothesised novel form of human-centred risk was tested and fully endorsed by the data. The evidence for these concepts was validated by an expert panel in NATS. An early form of heuristic notation designed to help decision makers conceptualise the content of those risks failed to be transferred for reasons which are discussed.

The second project (Hurisk 2) was a system development project. The aim of this was to provide a tangible tool to measure the "potential of safety attrition". This drew on data for the risk forms uncovered in Hurisk 1 case studies and expert workshops. Further interview and review work produced a set of working models for safety attrition and this create a "concept house" in the form of a software prototype, also called Hurisk.

This prototype were developed and tested via further NATS case studies. The results of these gave support to three sub formal, or heuristic, reasoning models. NATS experts were able to validate and use these in live project tests. Those models mapped to two domains, these were specific risks to projects and safe operations, which could be articulated as events, and a

bespoke concept which was called "strain". Measurement and visualisation concepts for these underwent further user trials and expert review. The proof of concept was established.

A number of soft concepts were tested to develop ideas for lesson learning approaches and tools. These were less successful, but their concepts remain of interest. A brief treatment of their design and the discussion of some validation data is given.

Hurisk tools were compared with the, largely engineering based, safety tools which were already in operation in NATS at that time was tested. A direct comparison between Hurisk and NATS' own TRACER, itself a decision monitoring device, was made. Hurisk was also compared with the requirements of a sample of NATS' own specific safety standards. This was done to give insight into the potential support a tool like Hurisk might give to their execution. These exercises, in management speak, helped create a "value proposition" for Hurisk. A basic description of the outcome of this process is given.

3.1.1 Some key points made in this chapter

- NATS operations managers found themselves to be arbiters of a new form of risk which they had no measurement system for.
- This new form of risk lay outside their considerable expertise in dealing with well defined human and technical risks.
- NATS found itself responding to 'modernity risk' in the form of increasing commercial pressures competing with a heretofore non negotiable level of safety.
- NATS was willing to look for an undetermined risk source in the form of 'reasoning deficiencies' which might be endemic in its management system, for example the long held belief in brainstorming as a good judgement process.
- It was not the significant or obvious risks which were the important object of NATS' attention in this project it was the many 'grains of sand' risk that might lead to an imperceptible attrition of safety.
- Initial attempts at a risk notation form for heuristic risk proved to be too far outside the reference grammar of the users and hence it was rejected.
- There was a change in reasoning focus between the two projects from risk identification to the management of risk knowledge.
- In the case of operational and engineering strain measurement a "journey system" i.e. one which took the user on a rational exploration of the complexities of the area was proved to be highly transferable even though it was a lengthy process.
- The use of simple weighting structures as an indication of preference was very easily achieved and produced an elegant form of comparison within and between users over time.
- The acceptability of the explicit use of a heuristic concept is shown in the "risk tolerance" measure which was purposely designed to operate emotionally and was accepted by users to do such.

- NATS had experimented already with heuristic measures for sensitive areas of risk, but these were deliberately not attempting to measure these risks.
- NATS documented safety requirements would, in many cases, be heavily supported by the introduction of a tool like Hurisk.

3.2 The National Air Traffic Services Project: Case study one

Introduction

An explorative research project on the attributes of human risk in air traffic control management decision making (Hurisk) was conducted using psychological field research tools. The research question surrounded the feasibility and acceptability of "non technical", "human centred" forms of risk measurement. The project was constructed from a number of smaller case studies. These were used to further develop and test a hypothesised descriptive model for the function of risk in air traffic (management) decisions.

Following investigative case studies a prototype tool was designed for use by decision makers. This introduced and tested a novel form of risk notation as a 'heuristic reasoning' device. An expert panel from the target population assessed the tool and its utility in the assessment of potential loss of safety through human judgemental processes. The notation was rejected by the users. The hypotheses however were supported.

3.2.1 Aims

The project had two aims:

1. Perform a series of representative case studies on current forms of risk-based decision making for key projects and map out key reasoning deficiencies.
2. Provide a new form of decision support for management and mitigation of risks.

3.2.2 Objectives

The project had three objectives

1. To articulate the 'human element' of risks and their associated decisions
2. To propose a risk-decision support notation tool

3. To assess the viability of the use of heuristic devices for assessing safety in relation to human decisions (to change the air traffic system).

3.2.3 Hypothesis development

Working closely with NATS experts on a steering committee three working hypotheses were formed. Two were developed from expert suggestions, the third was developed in a brief case study (not described).

1. Conventional safety tools measured "what safety tools measure" and that configuration left potential 'gaps' at the human behavioural interface.
2. A process of "safety attrition" could go undetected given the complex, interdependent and dynamic nature of modern air traffic control provision.

The third hypothesis comes from a simplified structure exemplifying NATS' three-fold mission to be safe, to meet the airline demand for service and to be efficient represented as a dynamic triad (figure 1). The key influences on the elements of this triad were made up of elements from a second order triad (figure 2)

Figure 1: SEC triad

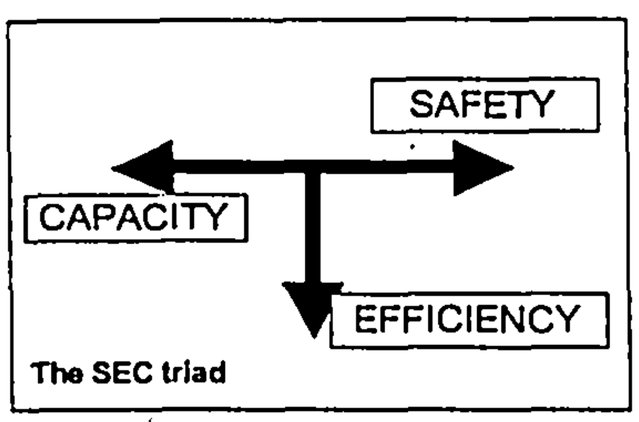
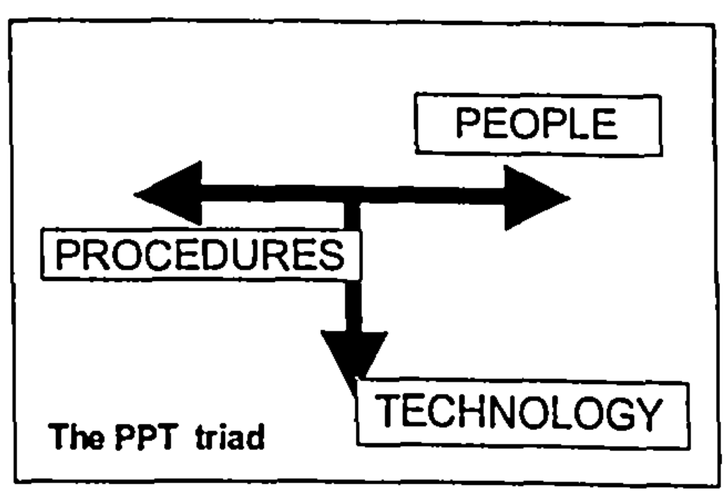


Figure 2: The PPT Triad



3. Triads hypothesis: Air traffic control system could be summarised in a simple triad structure in which safety, efficiency and capacity were related, and could be in competition. Within each arm of the triad a similar, second order, triad existed representing the relationship between decisions to change people, decisions to change procedures and decisions to change technologies.

3.2.4 Methodology

The research was conducted into two sections. Section one was a series of case studies.

Section two was two structured workshops. Techniques used in the case studies were:

- explorative and depth interviews
- observations (including observer participant mode when appropriate)
- group discussions
- in-house document review
- structured workshop

Table 1 overleaf shows a map of each of the smaller case studies linked to the strategic decision area it would exemplify.

Table 1: Case Studies in Hurisk 1

	Strategic Decisions affecting operating air traffic
Re-organisation of Air Traffic Control Training	There was a perceived need to re-organise the way training took place and this was being planned during the Hurisk project. The 'human risk' element of the training at that time was described.
Emergency Training in Air Traffic Control	In service Emergency Continuation Training (ECT) is a required element of every air traffic controllers career. Annually controllers must re-train to remain fit to deal with rare anomalies and emergencies which nonetheless do turn up. ECT requirement is an example of the risk averse nature of the ATC system as a whole.
Combined Control Function	In the 1990's air traffic volume growth was described as "phenomenal". A concept to integrate radar approach for Heathrow, Gatwick and Stansted was launched to deal with the fact that, given new traffic volumes, these three airports were too close together. This combined function would be called Terminal Control (TC). The relocation of the control function and its controllers was part of a larger planned strategy anticipating (stimulating?) growth at Stansted dating from the late eighties. The plan attempted to learn from the U.S. model with this problem which was to co-locate approach controllers and Terminal Manoeuvring Area outbound controllers with an underpinning assumption was that there would be capacity benefits. Co-ordinating of these functions was part of a larger five stage rationalisation plan for the airspace. Divergence from the predictions led to some of the scheme being "effectively scrapped". There were at the time calls for that whole ethos to be re-visited. A dispersed and evolutionary model was eventually adopted to cope with air traffic growth.
TCUPPER 'Resectorisation'	Observations were made over a day of a simulation trial for TC Upper resectorisation. Reduced vertical separation minimum (RVSM) took effect in the U.K. in Feb 2001. The changes to the terminal control upper sector came in response to the huge growth of Luton and Stansted. There was already a lot of congestion in terminal control north and delays in the London upper and London middle sectors were increasing.
2.5 Nautical Miles Separation	Previously the safe limit set on the distance between aircraft was three nautical miles(nm). This 3nm represents limitations of technology, primarily radar, and so called "wake vortex separation" i.e. the necessary distance between certain types of aircraft in sequence. These distances interact, depending on aircraft type and traffic configuration. For certain situations, this distance was been reduced to 2.5nm during complex (well understood) manoeuvring in the approach to land.
	Technology Innovations and Engineering Projects
Semi Automatic Meteorological Observation System	Cloud cover (height and type or mix), the visibility (runway and airport) and present weather, are key measurements which may change dramatically in an hour. At the time of the project all 13 UK airports had an assistant air traffic controller (AFTS a semi-clerical role). This person made weather observations every hour and reported these via computer data known as a METAR. These observations were partly automated e.g. remote temperature sensing. Cloud cover and the airport visibility are key information which require an observer. AFTS were trained by the MET office for this. The cost of employing these staff was high and other automation was shrinking their role. If the weather observations could be automated, a justification for abolishing the role would be high on financial reasoning.
Minimum Safe Altitude Warning	Controlled Flight Into Terrain (CFIT) is an important issue and the rate was still relatively high. These are where most fatalities occur. MSAW is a system which takes radar track information and detects aircraft proximity to terrain, or obstacles, and then alerts ATC in time to safely manoeuvre the plane. Designed as a safety net, the system catches incidents all other systems fail to. At the time the U.K. was considering committing to MSAW.
Final Approach Sequencing Tool	The Final Approach Sequencing Tool (FAST) was aimed at Heathrow where aircraft landed at a rate of 47 per hour during the busy periods on a tight application of existing rules and restrictions. Delays had the capacity to create large scale "ripple effects". Three controllers run four holding stacks at Heathrow in an involved arrangement. Final directions to the aircraft rely on "path stretching", and delicate control of turns married to standardised speed and headings. Controllers fine tune spacing with deceleration. This is sophisticated controlling unique, in the U.K., to Heathrow. Perturbations in the order would greatly increase controller workload. FAST was a semi-automation of some of these tasks and had been demonstrated in real time trials to show a reduction in some workload.

	General
The views of air traffic controllers	A series of one to one and group interviews with Air Traffic Controllers covering their views on the safety of air traffic and the "human element" of risk was conducted.

Testing the triad structure in case studies

A protocol in eight of these case studies was used to assess how the case study content area could be phrased within the grammar of the hypothesised triad elements. The following case studies were mapped to the triads:

3. 2.1 Case study one: Combined Control Function

3.2.2 Case study two: Re-organisation of Air Traffic Control training

3.2.3 Case study three: SAMOS

3.2.4 Case study four: FAST

3. 2.5 Case study five: MSAW

3.2.6 Case study six: 2.5 Nautical miles separation

3.2.7 Case study seven: Accident training in Air Traffic Control

3.2.8 Case study eight: The feasibility of safety attrition metrics, a workshop

These case studies were specifically chosen to map to the hypothesised interactions between the two triads. This is shown in the table 2 below.

Table 2: Mapping case studies to Triads

	People	**Technology**	**Procedures**	
Safety	3.2.7	3.2.5	3.2.6	3.2.8
Efficiency	3.2.2	3.2.3	3.2.1	
Capacity	3.2.1	3.2.4	3.2.1	
	3.2.8			

Validating the findings in workshops

The findings of the case studies and the triad notation results were formulated into a field report. A panel of NATS experts assessed this report to look for any "modelling value" from the case study conclusions. The panel was chosen from the target decision maker population. Operational air traffic controllers now also joined the review team. The review work was done in two structured workshops, these were:

1. Hurisk Concepts (preliminary findings review)
2. The Feasibility of Safety Attrition Metrics (a tool design workshop)

3.3 Results of case study one

The results, for the purposes of this thesis, will focus on

- The success of the notation technique
- The heuristic reasoning devices proposed
- Evaluation of the three hypotheses.

Treatment of case study results

These case studies, given their highly explorative nature, generated lengthy and detailed descriptive accounts of many the inner workings of the air traffic professions' decision taking. These narrative accounts were subject to a three stage content analysis (resolved by facilitated expert panel discussion). The stages were:

1. A coarse filter looking at whether a specific finding was indicative of a "barrier" to risk entering the system or an "inducer". This categorisation was guided by assessing the participant responses to two interview questions common to all case studies:

    a. What evidence was there to support the idea that *case study X* was presenting effective barriers to stem the flow of risks in the air traffic system?

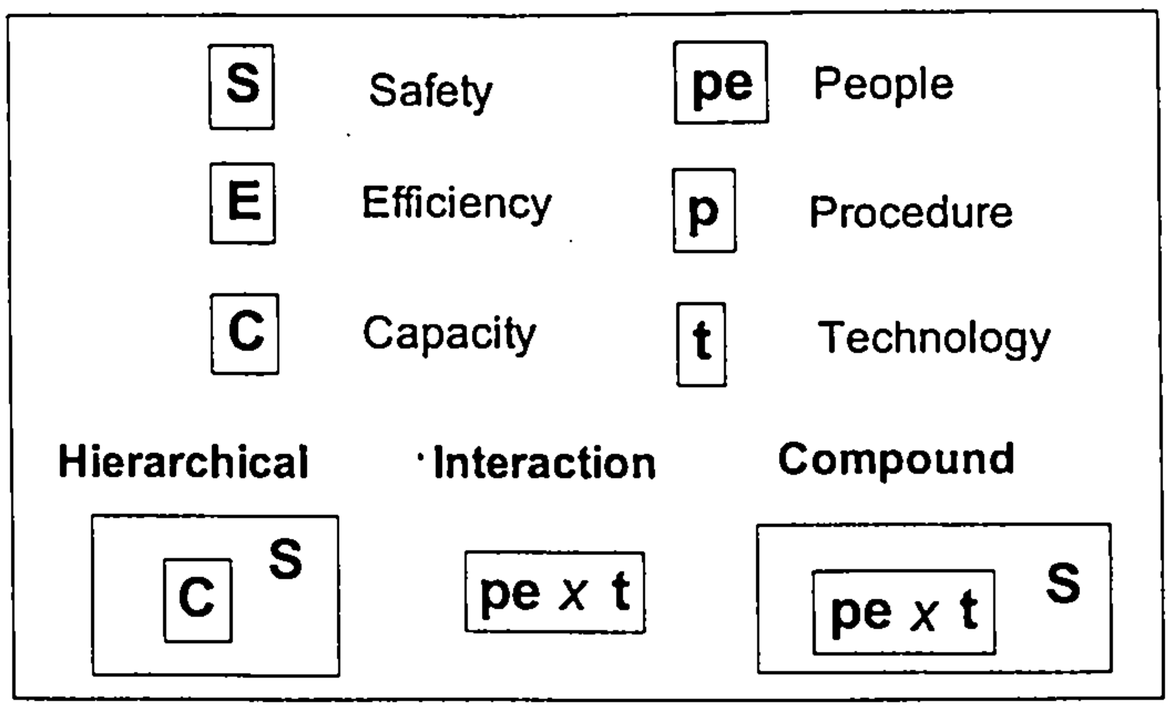
    b. What evidence was there to support the idea that *case study X* was potentially inducing risk within the air traffic system?

2. An assessment of the contents of the individual enablers and barriers fitted to the triad notation.
3. A combination exercise to assess the fit between the resultant elements and three "narrative schemes" underscoring hypothetical relationships between the triad elements.

Risk notation

These content analyses gave and overall risk notation heuristic. This scheme is shown in figure three below. The first order triads show Safety (S), Efficiency (E) and Capacity (C). The second order triads is the degree to which the risk relates *primarily* to People (Pe), Technology (T) or Procedures (P). Three possible relationships between these elements were given as: hierarchical; interaction; compound.

Fig 3 Triad notation



Examples of risk content analysis findings

It is informative to look at examples of barriers and inducers. A representative range of these taken from each case study result is shown in table three below.

Table 3: Examples of risk inducing and risk barriers

Case Study	Example Risk Barrier	Example risk inducer
Re-organisation of Air Traffic	The fact that ability spreading takes place is a risk strategy	The fact that the training is received from non operational personnel and that the

Control Training		live environment is highly dynamic may introduce risks but they would be very difficult to uncover and trace.
Emergency Training in Air Traffic Control	Air traffic controllers who fail their emergency and continuation training lose their certification for the site	The design and execution of emergency training at the units bears the burden of the risk of poor training design and poor or negative transfer of training.
Combined Control Function	Generalised problems with U.K. airspace were envisaged and action taken	Generalised problems with U.K. airspace were envisaged but impetus for large scale revision was lost
TCUPPER 'Resectorisation'	Constant recognition of factors which induce high workload	Acceptance of and acquiescence to growing congestion levels with no core questioning of the rationale
2.5 Nautical Miles Separation	We won' t take something we can't do safely" 2.5 Nm separation was seen as an example of where this happened and that was why all the additional criteria were added to mediate safety in its use.	controllers make their own safety analysis on when to use 2.5
SAMOS	The ideal is a 100% correct METAR. In an automatic system the risk is technology based. The error source is systematic not subjective.	it needs the specification of its parameters and these have to be managed in a safety critical way, it is run by software which will have to be implemented with the reliability issues this suggests and it is susceptible to common mode failures
MSAW	The system is designed as a safety net, it catches those incidents when all other systems fail to detect the proximity.	Technology induced role conflicts e.g. between pilots and controllers over height responsibility due to MSAW
FAST	Heathrow and Gatwick are recognised as near capacity	If FAST does not guarantee conflict free trajectories who is responsible for policing its advice?
The views of air traffic controllers	*Safety is felt to be eroding when it is no longer easy to do the comforting things e.g. winding the radar out to 60 miles (as opposed to standard 45) to take a look for upcoming traffic.*	*Controllers who operate a "climbing mentality" were felt to be misguided.*

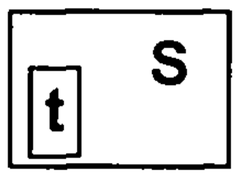
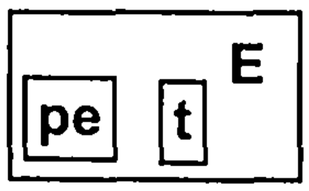
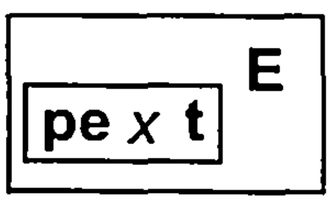
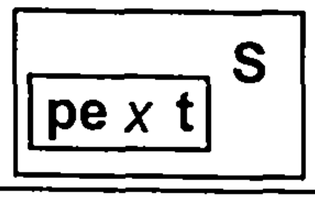
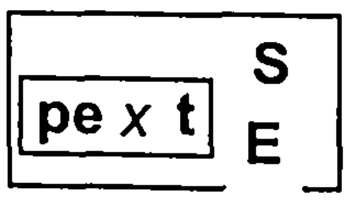
Examples of triad notation

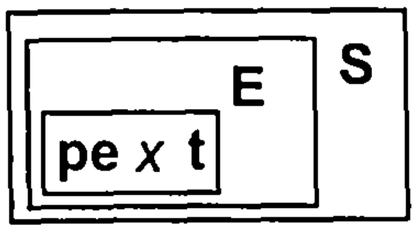
An example set of the results of the triad content exercise is shown in table four below. As well as the three triad element relationship schemes simple forms of barrier and inducer, i.e. where the barrier or inducer relates merely to one arm of the notation, were also observed.

Table 4: Examples from triad breakdown of risk inducement (first pass)

t	▪ Technology diversity issues e.g. black and white screens ▪ The introduction of "solely advisory" technologies with no testing on their actual use by the target users ▪ The knowledge risk associated with not knowing how to track the development of cutting edge technology applications e.g. neural nets in SAMOS ▪ If the forecast is not seen as reliable why is it factored in when reducing the required minimum separation? In wake vortex terms this is only permissible in favourable weather ▪ Maintaining a competitive edge through acquiring better technology than rival service providers is not a safety centric activity
p	▪ The interaction of new procedures with old procedures given familiarity index of air traffic behaviour as a help to good controlling
pe	▪ The threat of having your product (project) shot down by controllers is a difficult issue to factor in ▪ The lack of human factors centred training upgrades is problematic ▪ Those who never re-certificate but continue to train have a risk of skills degradation and skill shelf life ▪ The fact that the training is received from non operational personnel and that the live environment is highly dynamic may introduce risks (these would be very difficult to uncover and trace). ▪ The subject matter of training e.g. "decisiveness" has no apparent associated objective measurement criteria

pe x t	▪ The possible interactions between skill equalising tools ▪ *The need to avert low acceptance technology transfer failure leads to convoluted (and therefore potentially more political and risk inducing) commissioning routes* ▪ The use of human performance benchmarking (when it is poorly understood) for technology specification ▪ Assumptions of technology competence and presumptuous commissioning decisions which are made in "systemic isolation" ▪ The use of controller optimiser technology raises serious questions not only about who is the beneficiary but how stable the benefit is. ▪ Increasing the number of absolute error sources in automating a human function
pe x p	▪ The interaction of new procedures with old procedures given the lack of understanding of the implications cf software rusting as a parallel.
t E	▪ The lack of parsimonious integration of technologies e.g. FAST and STCA for engineering design expedience ▪ Lost expedience risk e.g. if SAMOS reduced the movements at an airport by one per hour the cost savings would be wiped from the project
t C	▪ Unintended side effects e.g. SAMOS measures weather and weather is factored into capacity so SAMOS reliability has a (non explicit) impact on capacity, e.g. MSAW may have runway shutdown side effects (capacity again)

	▪ Failure to be cautioned by existing systems functioning e.g. without FAST existing capacity gives a tight result using existing safety rules. This is an immediate cause for concern rather than a justification for a new piece of technology ▪ Safety critical parameter specification by non safety critical technologies e.g. SAMOS ▪ Technology which is not supposed to be for safety but wishes to impact efficiency e.g. SAMOS
	▪ The presupposition of a predictive relationship between research investment and technology implementation e.g. SAMOS, FAST. The pressure to commission is very great and may skew the analysis of the utility per se and the added value in the long term.
	▪ The credibility of advisory systems e.g. FAST does not guarantee conflict free trajectories. Who is responsible for policing its advice? ▪ Data structure weak links e.g. FAST depends on weather but the forecast at Heathrow is not reputed to be very good and decisions are taken to disregard it, how will the technology take such a decision? ▪ Planned shutdown in order to keep skill and awareness up would almost never happen at the busy periods which creates a risk hot spot.
	▪ Alert saturation mitigation lowering system functionality
	▪ General skill decay of the human base through automation

	▪ The evidence relating to recovery from technology failure e.g. if FAST was to go down the stacks lose their pattern for predictability with an adversely effect on workload even for experienced controllers.

3.4 Discussion of results

In modern air transport airport pressures primarily surround capacity and efficiency. Whilst meeting responsibilities to deliver the 'safe and expeditious movement of aircraft' NATS has a duty to assess every change in their systems with respect to safety. Whereas historically "the system" referred to the technological system (mediated by air traffic controller behaviour), this project was an attempt to expand that notion to the management decision making system also.

These case studies considered in conjunction with the outcomes two other risk programmes (Atcon and Softcon[2]), were an attempt to clarify how to support decision makers faced with increasing pressures in the real world of burgeoning air travel. NATS hoped this clarification would bring about concepts for tools to manage that risk.

The dominant tool in NATS safety management, at the non technical level, was the "safety brainstorming meeting". This is a semi formal technique where professionals explore safety implications of changes to the air traffic systems. These assessments were observed to be based an informal kind of optimisation based around consensual compromise. Risk measurement was observed to feature in this process, but it was little more than discussion concluded by a simple high, medium and low ranking scheme in a consensus process.

[2] Atcon and Softcon were major re-organisation projects in NATS looking at air traffic controller organisation and the new air traffic control systems and computer tools

A more formal tool, such as that which could be derived from the investigation into these triads, could be used to assess the nature of the "decision spaces" around these changes and developments. Such a tool could create some kind of enhanced structure for reasoning within which a primary regard for air traffic safety would always be maintained. Explicitly verbalising and developing management plans framed in the relationship between the two hypothetical risk triads, it was argued, could have been an important source of improved safety rationality. This would preserve but "tighten" the sub-formal reasoning which already took place. Certainly, it could cause a richer form of safety discussion than brainstorming risks for a higher order proposition and assigning one of three scores to them.

Having an improved structure to managers' thinking, like the one the triad model was proposing, could cover off the sources of human risk thrown up by management decision making. This was hypothesised by the expert panel as an important blind spot. These risks, if they were there, would currently be invisible to mainstream quantitative assessment techniques. Decision making and decision interpretation of human operators interacting with the content of the triads above could transform into an important, but usable, schema to focus risk measurement and control.

It was clear that this research was exploring a tension between the three parts of NATS mission (safety, efficiency, capacity) These could be seen as contradictory, for example large increases in airport capacity can dramatically affect an airport's safety in a negative direction. Furthermore failure modes in one area were highly likely to generate failures in the others.

The development of the interacting triads represents a heuristic simplification of the dynamic "socio-technical system" which the NATS experts were describing. Their intended use of the

triads notation was to exemplify key problem spaces. For example, the development of technology is vital for increased capacity. Fitting that new technology however is a safety critical exercise. The equipment must be highly reliable, highly usable, highly compatible and have an "at least zero" impact on safety. Likewise changes to procedures and human factors, e.g. the use of training and simulation, are also safety critical even if they are essentially aimed at efficiency or capacity.

Hypotheses one and two in the study were generally supported by the descriptive data. Significant gaps in risk reasoning were agreed to be present at the human decision interface. Any one risk thus caused, if it was considered in isolation, would seem small. However over the case studies as a whole there were large numbers of them. This volume issue would support the notion of a process of "attrition".

The triad notation applied to the "risk inducers" (as these were a key worry), attempted to ascertain two things.

- Could this notation create a decision short-hand to differentiate and cluster risk types meaningfully?
- Would this form of summary reasoning be accepted by air traffic control decision makers?

As has been demonstrated in a range of examples above, the triad structure did work descriptively. Perhaps it was at its weakest across such diverse case studies in being able to identify any "pure" procedural problems. Only one real example was uncovered. This might lead us to believe that the idea of procedural problems is not well enough defined. There are two alternative explanations, one is that the procedural area of NATS is so strong and stable the only risks it produces are in its own evolution. The second is that procedural risks are

being expressed better in the other levels of the triads. If so the legitimacy of a procedure arm of equal importance to the others seems weak.

Where the triads give their strongest indication of predictive power is clearly when they are anything to do with people. The majority of the stand alone and interactive risks surround people. This is perhaps not surprising, but again a model where equal weight might be given to a people arm would seem to undermine its explanatory value for risk. That said as a risk highlighting mechanism it would of course serve an important purpose.

Why didn't this notation system work?

The results of the expert review into the findings were that the triad notation would not go forward as a working concept. These early heuristic conceptualisations of risk were simply not informative or compelling enough to be seen as valuable to aid decision making. They did not differentiate risk magnitude when linked back to air traffic safety. A first major failure therefore was that they just didn't say enough comparatively. Whilst being relatively good at unearthing the complexity of problems, they were not perceived to "add any value" over normal narrative forms of discussion or reasoning. In short they did not really aid prescriptive rationality.

A second reason for the failure was that they were simply too alien to air traffic control decision makers. The metaphor, although comprehensible, proved in a workshop with NATS experts that it did not gel with their other forms of description or reasoning. The notation left people rather cold. Something better, and importantly, something closer to the home reasoning forms which people were already using was required.

3.5 Conclusions and further research direction

Although Hurisk missed the mark in terms of offering a new risk reasoning tool, the sponsors recognised with certainty that it hit the mark dead centre when it came to describing risk exposure that had never been described before. The long-hand forms of describing the risks posed by these technological and organisational case studies caused quite a stir in the organisation[3]. This is precisely because they were recognised for their validity.

The key finding of Hurisk was that it was indeed possible to articulate risk types that were below the radar of normal safety systems. NATS safety professionals had agreed that this worried them. Three conclusions were important for this thesis.

1. The important risks almost all came from new projects and the strain these added to previously secured safe working arrangements.
2. These risks were heavily biased towards "the human judgement" element of the projects, not the technical.
3. A way of summarising and scaling these risks was proven to be highly desirable.

The Hurisk 1 triads were based on an a priori assumption that safety, efficiency and capacity were in some sort of fixed relationship structure i.e. you couldn't change one without effecting the others. This assumption extended to the idea that people, procedure and technology considerations were in a similar relationship. Whilst these relationships might be parsimonious, measurement of their effects had not proved viable.

A second project (Hurisk 2) project was commissioned to focus on a single hypothesis. There was a source of risk, which is invisible to mainstream qualitative assessment techniques,

[3] The field report used for the two workshops was a collection of all the main findings from the work so far. When a senior manager in NATS saw a copy, his initial reaction was to insist that every copy be accounted for immediately and confiscated. This done, the same manager, after some further thought, photocopied the entire report to send to his seniors.

found in the decision making and decision interpretation of human operators and decision makers. It was agreed by the NATS expert panel that the Hurisk one concepts, if developed, might support decision makers to use their expert judgement to mitigate these risks. The result could be an increased sensitivity to air traffic safety from a human perspective.

'Concept clean up' and further in-depth research was commissioned around three assertions:

1. The air traffic control system was shown through Hurisk one to be under "strain" and it was desirable for that strain to be formally measured and controlled.

2. Risk assessment within NATS would benefit from and expansion of its definition to take into account those sorts of risks thrown up in Hurisk, and these should be formally measured and controlled in the future.

3. NATS was still not good at learning lessons from the past.

The aim of a second Hurisk project would be to deliver a heuristic 'reasoning system', this time as a computer-based risk and decision support prototype, to :

- Measure 'operational strain' caused by air traffic engineering projects
- Measure/compare risk posed by "the human element" of air traffic control project decision making
- Create usable, sub-formal, lesson learning capability

Provide a raft of "culturally acceptable" visualisation environments to support on-line reasoning.

3.6 The National Air Traffic Services Project: Case study two

Introduction

Detailed evidence for new sources of risk in air traffic control had been produced in an explorative project (Hurisk 1). This evidence, validated by an expert panel inside of NATS, highlighted that the decision making processes and decision maker behaviours were a key mediator of an emergent form of operational risk brought on by increasing commercial pressure. This risk was invisible to current mainstream qualitative assessment techniques.

Modelling from the evidence base of Hurisk 1, a detection and monitoring system for this new decision making risk was designed. This system utilised novel psychological modelling approaches and developed software prototypes to test these. A tool to support decision makers mitigate risks was the main outcome.

The results of the interviews, modelling methods and bench-testing work which will be referenced in this case study are not being reported as a part of this thesis. The reason for this is that, interesting though this research is, it is not the primary focus. Also, a consideration of those findings would be extremely lengthy and involved.

The discussion below is primarily a description of the system and not the wide research base used to create it. The following sections will focus therefore on some of the technical outcomes which are pertinent to the decision support concepts being explored later in this thesis.

3.6.1 Aims

Hurisk 2 focussed on types of risk introduced into the air traffic system in two main areas.

- "Strain" or perturbations that new air traffic control projects added to previously secured safe working arrangements.
- The introduction of risk through "the human judgement" element of these projects.

The task of the project was to investigate a useful way of measuring, scaling and communicating these risks. The aim of Hurisk 2 would therefore be to deliver a heuristic 'reasoning system', in a computer-based risk and decision support model, to:

1. Measure 'operational strain' caused by air traffic projects
2. Measure/compare risk posed by "the human element" of air traffic control project decision making
3. Provide a raft of "culturally acceptable" visualisation environments to support better risk reasoning
4. Create usable, sub-formal, lesson learning capability.

3.6.2 Objectives

Hurisk 2 aimed to develop a 'risk knowledge management system' rather than a risk identification system. Such a reasoning tool would have to augment the rationality, transparency and intuitiveness of group decisions on projects. A software system was considered to have the highest chance of success. To create the highest resonance for air traffic control a suitable metaphor was needed. The two strongest metaphors in that culture are gauges and checklists. The project undertook:

1. the development and testing of "strain gauges" to integrate existing strain style data into thinking
2. the development and testing of "risk gauges" to integrate risk data into decision making
3. the development and testing of a checklist based "lessons tool" to fold forward lessons learned

4. the development of one overall interface to house the above concepts

This large project was sub-divided into a number of systems development objectives. Table five below gives a summary of these.

Table 5: Summary of Hurisk 2 objectives

Strain Gauges	- Investigate feasibility of enhancing (by some form of aggregation or comparison) engineered measures in existence e.g. proximity, Safety Significant Events - Look for 'factors' associated with controllers and their managers which predicate the above - Comparison for parallels with existing incident investigation tool (TRACER) - Find a suitable case study to test concepts - Define user groups and needs through interview and workshop
Risk Gauges	- Develop a methodology to elicit key risks - Develop a usable and acceptable measurement form for risk - Find suitable case study to test concepts - Define user groups and needs through interview and workshop
Lesson learning	- Describe a lesson learning tool built upon: - Interviews with decision makers and potential users. - Adaptation of strain and risk data. - Reviewing existing projects for potentially generic material. - Define different user groups and needs for such a tool at strategic and operational level
System development	- Develop the mathematical underpinnings of the models - Develop culturally valid look and feel prototypes - Work with software provider to build system prototype - Test system in real test case - Generally assess Human Computer Interaction profile of system using verbal protocol technique
Document system	Working within a software development metaphor provide - System Architecture - Functional Specification - User scenario
Describe dissemination plan	- Understand the factors affecting the dissemination of such a tool within NATS

3.6.3 Methodology

Hurisk 2 was a larger and much more intuitive development project than Hurisk 1. Hurisk 2 two used seven methods to develop and test a prototype:

1. Working with the results from Hurisk 1 and combining these with new and extant risk measurement concepts.
2. Observation and interviews with decision maker groups (management and engineering) to develop strain measurement concepts and prototypes.
3. Modelling with past projects (management and engineering) to fit and refine Hurisk measurement concepts.

4. Interviews and workshops with potential end users to test acceptability of measurement and visualisation concepts.
5. Bench-testing of prototype tool in controlled trials with end users (management and engineering).
6. The development of suitable algorithms to reflect the heuristics embedded in the systems (e.g. risk tolerance, strain normalisation).
7. Benchmarking the final tool against an existing NATS tool and NATS safety standard requirements

3.7 Consideration of results

The results of Hurisk 2 development will be explained and then discussed in the following areas:

1. Strain scoring and adjustment
2. Risk scoring and adjustment
3. Risk and strain measurement tools and visualisations
4. Lesson learning measurement ideas and visualisation

3.7.1 Strain scoring and adjustment

From interviews and modelling of past projects it was clear that a questionnaire tool would be the final form of any capture mechanism for a strain concept. The final structure of the categories and questions (all scored on a scale from 1-100) will be discussed in section three. The questionnaires were designed in two forms, a short and a long. The algorithm for the final strain score (expressed as 1-100) in the short form was:

$$Short\ mean = \frac{1}{n}\sum_{i=1}^{n} S_i$$

Using category weighting

In the long version of the questionnaires the categories can be weighted in a number of ways so that the usefulness of a weightings structure to reflect preferences could be explored. The examples below show weighting data was used to aid understanding. These were:

1. Rank order comparison of high level category weighting choices between organisations (table 6)
2. A two tier weighting method allocating weights to lower and higher order categories (table 7)
3. Comparison of weighting between trials (shown as a delta value) (table 8)
4. A comparison of weighting between individuals (table 9)

These tables below shows actual weighting patterns from the development case studies.

Table 6: Mean weighting structure ATC centres 1 (3 participants) &2 (6 participants)

Weighting order study 1	mean
Controller working patterns	87
Uncertainty from decision making	85
Sressors	78
Controller role and function	73
Decision attributes	73
Perceptions of stressors	72
Stability of relationships	63
Decision/project planning	57
Controller skill impacts	57
Decision consultation issues	42
Perceptions of decision making	42

Weighting order study 2	mean
Controller working patterns	83
Controller skill impacts	82
Perceptions of stressors	79
Controller role and function	78
Uncertainty from decision making	73
Decision attributes	72
Perceptions of decision making	72
Decision consultation issues	71
Stability of relationships	71
Sressors	70
Decision/project planning	68

A two tier weighting pattern was experimented with, weighting high level categories and main categories. This is shown in the table below.

Table 7: Comparison of two tier weights

	participant 1	**participant 2**
Human Factors	0.26	0.20
Controller skill	0.1	na
Controller role and function	0.5	0.5
Controller working patterns	0.4	0.5
Planning	0.26	0.50
Decision project planning	0.5	0.3
Consultation issues	0.5	0.7
Stability and stress	0.16	0.20
Relationships	0.2	0.333
Stressors	0.2	0.333
Perceptions	0.6	0.333
Decision making	0.32	0.10
Decision attributes	0.2	0.25
Perceptions	0.2	0.25
Uncertainty	0.6	0.5

Weights were also used to highlight changes in priority within participants. The table below shows weighting patterns used to compare within a subject over time in an engineering strain project.

Table 8: Comparison of shifts in priority for strain categories

Strain category	T1	T2	Diff
Staff Roles and Responsibilities	1	1	0
Managing re-specification	2	2	0
Re-specification	3	6	-3
Reference points	4	14	-10
Time management	5	3	2
Consultation	6	5	1
Politics and Management	7	11	-4
Pressure and uncertainty	8	9	-1
User "preferences"	9	7	2
Complexity	10	12	-2
Specification	11	4	7
Non standard processes and technology	12	13	-1
Forces in equipment change	13	15	-2
Time and technology	14	8	6
State of affairs	15	10	5

Weighting patterns were also used to make rank order comparisons between users on the same project as the comparison between an engineer manager and a project engineer on the same project.

Table 9: Weighting differential between users

Attributes	Proj man	Engineer	diff
Staff Roles and Responsibilities	1	1	0
Managing re-specification	2	9	-7
Re-specification	3	7	-4
Reference points	4	8	-4
Time management	5	3	2
Consultation	6	4	2
Politics and Management	7	5	2
Pressure and uncertainty	8	2	6
User "preferences"	9	6	3
Complexity	10	10	0
Specification	11	11	0
Non standard processes and technology	12	13	-1
Forces in equipment change	13	15	-2
Time and technology	14	14	0
State of affairs	15	12	3

Individual strain score weighting calibration

Due to the heuristic nature of strains (see discussion) a simple mechanism to normalise the contribution of any one strain to the overall score was sought.

Simple strain scoring

Hurisk returns a grand score in the scale of 1-100 for strain. The simplest example of this is where the short forms of the checklists are used. Here all the raw scores are simply summed and divided by the number of questions.

Weighted strain scoring

In the case of the long versions of the checklists users are permitted to add subjective importance weights to individual categories. These are also scores from 1-100. This presents the need to translate both the weights and the resultant data to preserve the scale. In the case of the weights the transformation is a simple one, all weights must sum to one. Thus, the weights act in an unsurprising way to modify relative contribution.

In the case of the resultant data the final mean score has to be multiplied by the number of weight transformations, in this case, the number of categories. This has the effect of returning the final score into the same scale as the raw scores, i.e. from 1-100. The final mean is now affected by a relative contribution of scores based on weightings but remains in the same scale. Figure six below shows a simplified worked example of this principle.

Figure 4: Weighted means explained

Weights are given as a score out of 100

Then translated to sum to one

Category	Raw scores	Category Weights	Actual Weights	Weighted Scores
Cat1	88.00 77.00	80	0.44	39.11 34.22
Cat 2	20.00 20.00	60	0.33	6.67 6.67
Cat 3	45.00 45.00	40	0.22	10.00 10.00
Mean score	49.17		Mean score	17.78
			Adjusted Mean Score	53.33

The mean has to be adjusted to keep the final score in the same scale as the original. The weights effect relative contribution to the mean

Category	Raw scores	Category Weights	Actual Weights	Weighted Scores
Cat1	88.00 77.00	40	0.22	19.56 17.11
Cat 2	20.00 20.00	80	0.44	8.89 8.89
Cat 3	45.00 45.00	60	0.33	15.00 15.00
Mean score	49.17		Mean score	14.07
			Adjusted Mean Score	42.22

In this case the larger weighting on the lower scoring category has given a lower overall mean

A possible problem

In figure one above the system explained works, but it has a simple weakness, all of the raw data is contained in categories of equal size. All strains in Hurisk are designed to be of equal importance before the user weights are added. However the Hurisk categories do not contain equal numbers of strains. Using the method in figure one above would give the following problem:

Strain X is in category Y. Y contains 13 strains in total. A score for strain X will contribute one over thirteen of the strain derived from category Y.

Strain B is in category W. B contains 5 strains in total. A score for strain B will contribute one over five of the strain derived from category W.

Strain category scores Y and W will, in turn, be averaged into the final strain score.

Simply put, strain X has less "power" than strain B to effect this final strain score. Strain X, by default, is less important than B. As all strains in Hurisk have to be of equal importance before user weights are added. A simple solution to this problem is needed.

A simple solution

The simple solution might be to generate a final strain score not as a grand mean expression of each category mean but as a grand mean of the total number of checklist items once they have been weighted. Thus all the weighted scores are treated as being in one large category for the purposes of generating the final strain score. This solution, however, may make the weighting and scoring process a little less intuitive. It also has the disadvantage of not allowing the mean scores for categories to be examined in the fish eyed lens principle at a level in the hierarchy one above that of individual strains.

A more elegant solution: balance weights

To get around the above problem each strain can be given a balance weight. This weight is a reflection of the size of family to which the strain belongs. This means that strains from large families are given a mathematical boost. This boost is precisely relative to the size of the family compared with the total number of strains in the checklists. Conversely strains from small families are suppressed slightly relative to the size of the family compared with the total number of strains in the checklist.

For example, imagine a total checklist made up of 15 items with the scores shown in table one. A problem occurs when mean scores derived from families of unequal size are combined to form a grand mean. Compare case 1 and case 2, in case one the score of 50 is increased by 10 (shown by the bold figure in column three), in case two likewise a score of 50 is increased by 10. In case 1 double the effect on the grand mean is observed for the same unit change in one strain only. Effectively changes to the strain in the smaller family are twice as effective.

Table 10: Illustrating the need for balance weights

		CASE 1		CASE 2	
cat 1	**cat 2**	**cat 1**	**Cat 2**	**cat 1**	**cat 2**
100.00	100.00	100.00	100.00	100.00	100.00
50.00	100.00	**60.00**	100.00	50.00	100.00
60.00	50.00	60.00	50.00	60.00	50.00
70.00	50.00	70.00	50.00	70.00	**60.00**
80.00	60.00	80.00	60.00	80.00	60.00
	60.00		60.00		60.00
	70.00		70.00		70.00
	70.00		70.00		70.00
	80.00		80.00		80.00
	80.00		80.00		80.00
Mean	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**
72.00	72.00	74.00	72.00	72.00	73.00
Grand mean	72.00	**Grand mean**	**73.00**	**Grand mean**	**72.50**

Table 11: Balance weights based on family size

Adjustment					
X 5/15	X 10/15	CASE 1		CASE 2	
cat 1	**cat 2**	**cat 1**	**Cat 2**	**cat 1**	**cat 2**
33.33	66.67	33.33	66.67	33.33	66.67
16.67	66.67	**20.00**	66.67	16.67	66.67
20.00	33.33	20.00	33.33	20.00	33.33
23.33	33.33	23.33	33.33	23.33	**40.00**
26.67	40.00	26.67	40.00	26.67	40.00
	40.00		40.00		40.00
	46.67		46.67		46.67
	46.67		46.67		46.67
	53.33		53.33		53.33
	53.33		53.33		53.33
Mean	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**
24.00	48.00	24.67	48.00	24.00	48.67
Grand mean	36.00	**Grand mean**	**36.33**	**Grand mean**	**36.33**
Adj Grand mean	72.00	**Adj Grand mean**	**72.67**	**Adj Grand mean**	**72.67**

In the design principles of Hurisk the family to which a strain belongs is serendipitous not deterministic. Thus all strains must be equal. Table two shows the effect of balance weighting. The scores are multiplied by a factor based on the group number (5/15 or 10/15) and the resultant means have to be adjusted, as in the previous example, to re-scale them.

The solution

Strains from the small family are suppressed and strains from the large are boosted relative to the total number of items (in this case 15) and a unit change in the strain value is not sensitive to the number of strains in the category. The user does not see this algorithm at work it is hard wired into the calculations. Importantly if new strains were added or different strains used, this algorithm would require re-calibration. The means are given by the following:

$$Category\ Mean = \mu_j^C = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^S * S_{ij}$$

$$Overall\ Mean = \mu = \frac{1}{m}\sum_{j=1}^{m} w_j^C * \mu_j^C$$

3.7.2 Risk scoring and adjustment

Risk events in the system are scored, as strain events are, on a scale of 1-100. The system returns a final single risk score in this range for every risk. The calculation of this risk score can involve a seven part system. The various scoring options are described.

One component scoring

A single overall risk score can be given a score in the form of 'high' (returns a value of 75), 'medium' (returns a value of 50) or 'low' (returns a value of 25). Alternatively this overall risk score could be given a discrete value from 1 to 100.

Two component scoring

The overall risk score can be made up of two components. These can be scored in exactly the same way and their scores combined to give the overall score as a function of their values.

Six component scoring

Each of the two component scores can be made up of three other scores using the same principle above. These, lowest level, scores propagate the two level scores above which in turn propagate the single overall score. These scores relate to

- Probability (P)
- Control (C)
- Effect (E)

- Knowledge of probability (kP)
- Knowledge of control (kC)
- Knowledge of effect (kE)

Risk tolerance scoring

The two component scores are combined to make overall score by a rule which is governed by an algorithm. This user controls the application of this rule by applying a score of 1-100 to indicate their risk tolerance. This variable sets the sensitivity of the algorithm.

How the risk score is calculated

$$\text{Risk} \quad R = \sqrt[3]{P * C * E}$$

$$\text{Confidence } K = \sqrt[3]{kP * kC * kE}$$

$$\text{Adjusted Risk } A = \min\left\{R + X\left(\frac{100-K}{2}\right), 100\right\}$$

Adjusting Risk: risk tolerance explained in lay terms

The combination of confidence and risk scores is mediated as discussed above by a score for risk tolerance. Take the example of a risk score of 50 with an associated confidence score of 100. In such a case Hurisk would be happy to return a score of 50. If the confidence was 50 however, things become more complex since this score reflects the fact that the user is uncertain if their risk estimate is correct. It could, in fact be lower, it could be higher. Hurisk adjusts the risk summary in three steps:

1. The distance between confidence and 100 is calculated and divided by two. This represents the uncertainty that the score could be larger or smaller and attributes half the difference to each possibility.

2. The score from Step 1 is weighted by the risk tolerance score (to increase or decrease the effect)

3. the weighted score is added to the risk score and compared to 100, the minimum score is used.

Risk distance plotting

The distance between two risk users' scores (confidence K and risk R) is plotted in one of the system's visualisations (see description below). The coordinates are given by:

$$\text{RiskDis } D = \sqrt[3]{(P2 - P1)^2 + (C2 - C1)^2 + (E1 - E2)^2}$$

$$\text{ConfDis } B = \sqrt[3]{(kP2 - kP1)^2 + (kC2 - kC1)^2 + (kE1 - kE2)^2}$$

The calculation of a decay function in the audit bar

The system contains a visualisation to express "depth of analysis". One, if users have used the minimum scoring methods this would return a low score. To reflect the shelf life of scores (however generated) a decay function begins. This is given by the explanation overleaf:

Decay $Z_t = Z_{t-1} + \varepsilon_t$ where the perturbations ε_t, $t = 1, 2, 3, ..$are all independent and normally distributed $\varepsilon_t \backsim N(0, v_t^2)$

{The last equation reads that the measurement Z_t taken at time t is the measurement Z_{t-1} taken at the previous time point $t-1$ plus a normally distributed error ε_t having a zero mean and a known variance v_t^2. It id often appropriate to assume that v_t^2 does not depend on the time t of the obsewrvation. This hypothesis is sometimes stated as saying that the sequence $\{Z_t : t \geq 0\}$ is a "random walk". This assumption is widely used in statistics and probability and is one of the simplest ways of saying that the current measurement is similar to what it was previously but potentially perturbed up or down. The larger the variance the bigger the potential perturbation. The zero mean assumption ensures that an upward perturbation is equally likely to a downward perburbation.The final equation demonstrates that under the random walk hypothesis the longer we wait the more different the current measurement Z_{t+k} $k+1$ time steps ahead of the measurement we last observed Z_{t-1} might be. Thus this should read}

$Z_{t+k} = Z_{t-1} + \varepsilon_t + \varepsilon_{t+1} + \cdots + \varepsilon_{t+k} = Z_{t-1} + \varepsilon_t(k)$
where $\varepsilon_t(k) \backsim N(0, v_t^2(k))$ and where the variance $v_t^2(k) = v_t^2 + v_{t+1}^2 + \ldots + v_{t+k}^2$.

{ This can be seen as a logical consequence of the random walk hypothesis and can be calculated by simple substitution. For example for $k = 1$

$$Z_{t+1} = Z_t + \varepsilon_{t+1}$$

but we know that

$$Z_t = Z_{t-1} + \varepsilon_t$$

so substituting for Z_t in the equaton above gives

$$Z_{t+1} = Z_{t-1} + \varepsilon_t + \varepsilon_{t+1}$$

Note that when $v_t^2 = v^2$ for all time then $v_t^2(k) = (k+1)v^2$ so that the variance of the perturbation increases linearly with the interval of time between the two observations.}

3.7.3 Risk and strain measurement tools and visualisations

Strain measurement and visualisation

The strains measurement tool took the form of a questionnaire. The questions were developed from a content analysis of Hurisk 1 data coupled to new data from interviews with engineering and management users and from modelling with past examples of projects. The questionnaires came in different forms.

The Hurisk "fish-eyed lens"[4] principle applies to strain checklists and risks similarly. There are two forms of strain questionnaire, a short and a long version. The short version is typically 20 – 25 items, the long 80-90.

Below is an example of a section of the short version of engineering strain. This relates to the category of "complexity and demands". Scorers were asked to provide a percentage value for the degree to which this strain is "present" in the project in question (previous scores shown in brackets)

Figure 5: Example from the strain tool

Short, Engineering strain questions
Complexity and demands category (1/8)

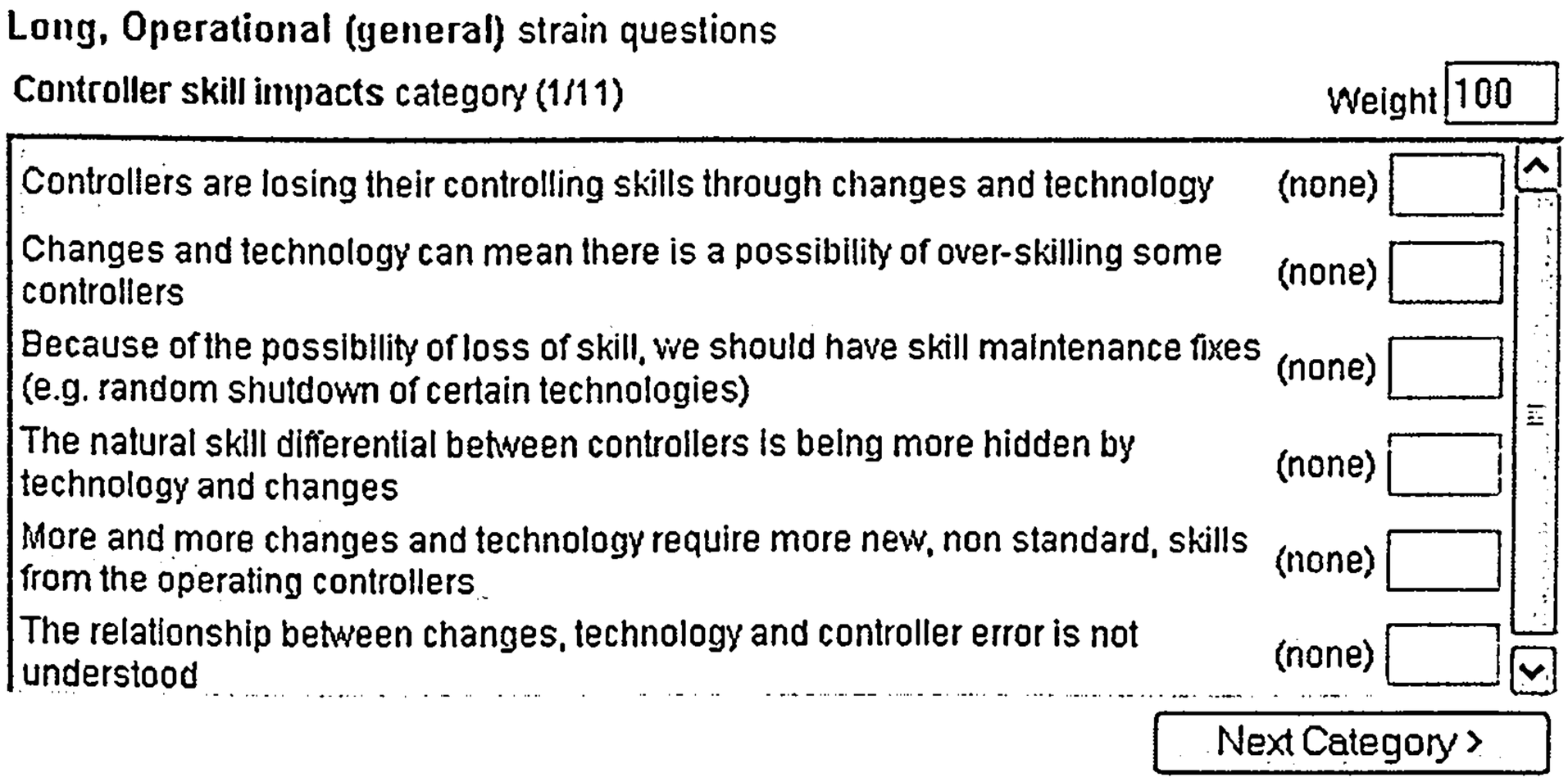
Key decision makers have many differences on a variety of strain issues for this project	(20)	
This is a large and or complex project	(45)	
This project has an effect on many other systems	(60)	
This project is addressing a system or state of affairs which has been under a lot of strain for a long time	(70)	

Next Category >

[4] The fish eyed lens is the way in which all data in Hurisk can be scored and viewed either at the very high level, often one score, or the very detailed level. Whichever level is chosen the scores are propagated up and down the system.

Below is an example of a category (controller skills impacts) from the long version of the generic operational questionnaire. Note a key difference that the entire category (one of eleven categories of questions) is allowed to be independently weighted. This questionnaire is used as a general health check on the operation in question.

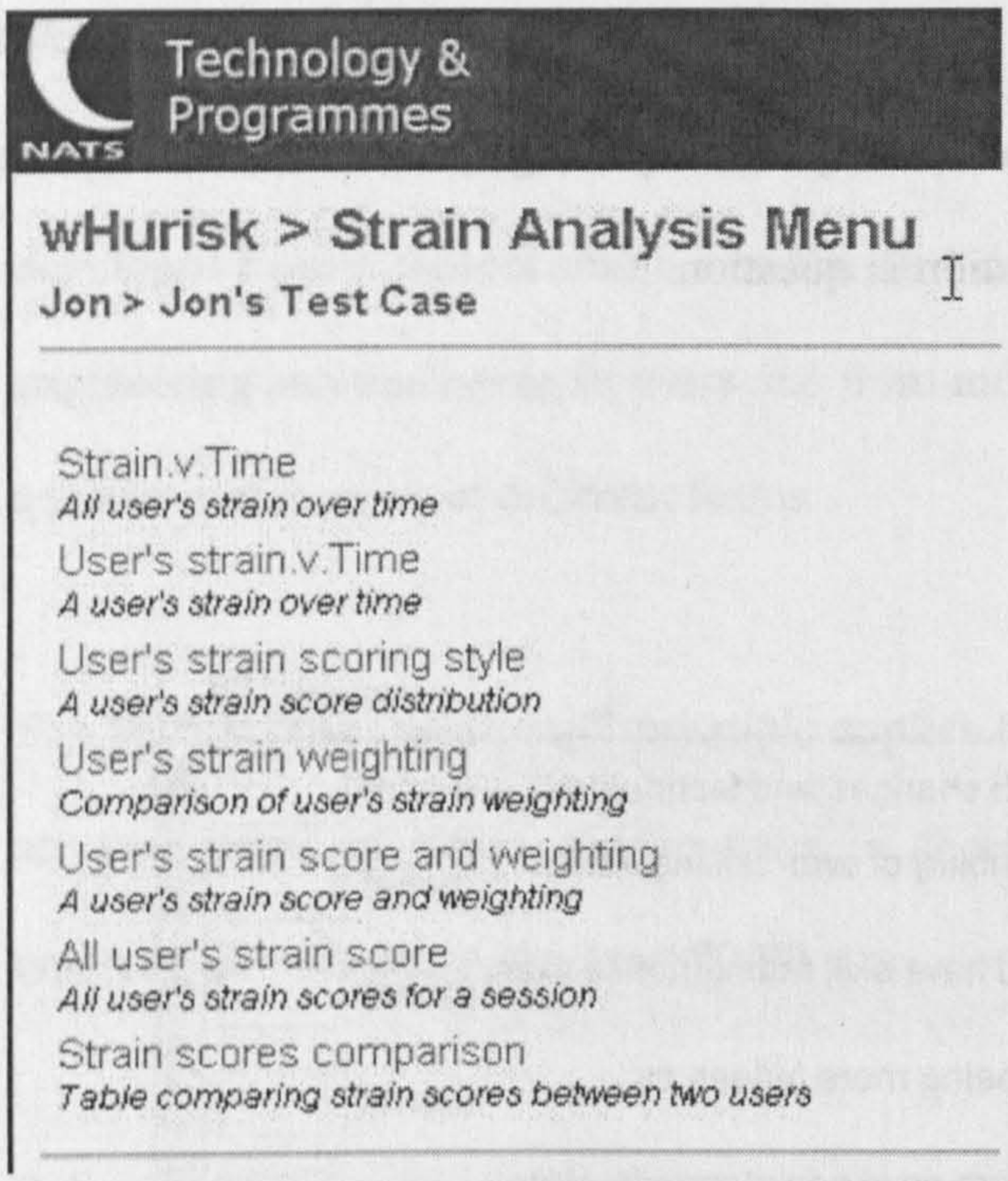
Figure 6: Example two from strain tool



Strain Visualisation

A key benefit of the Hurisk design is not only to provide an environment for structured scoring but also to provide a visualisation suite for reasoning. Strain can be visualised as we have seen in list form with attendant scores. There are also a number of graphics:

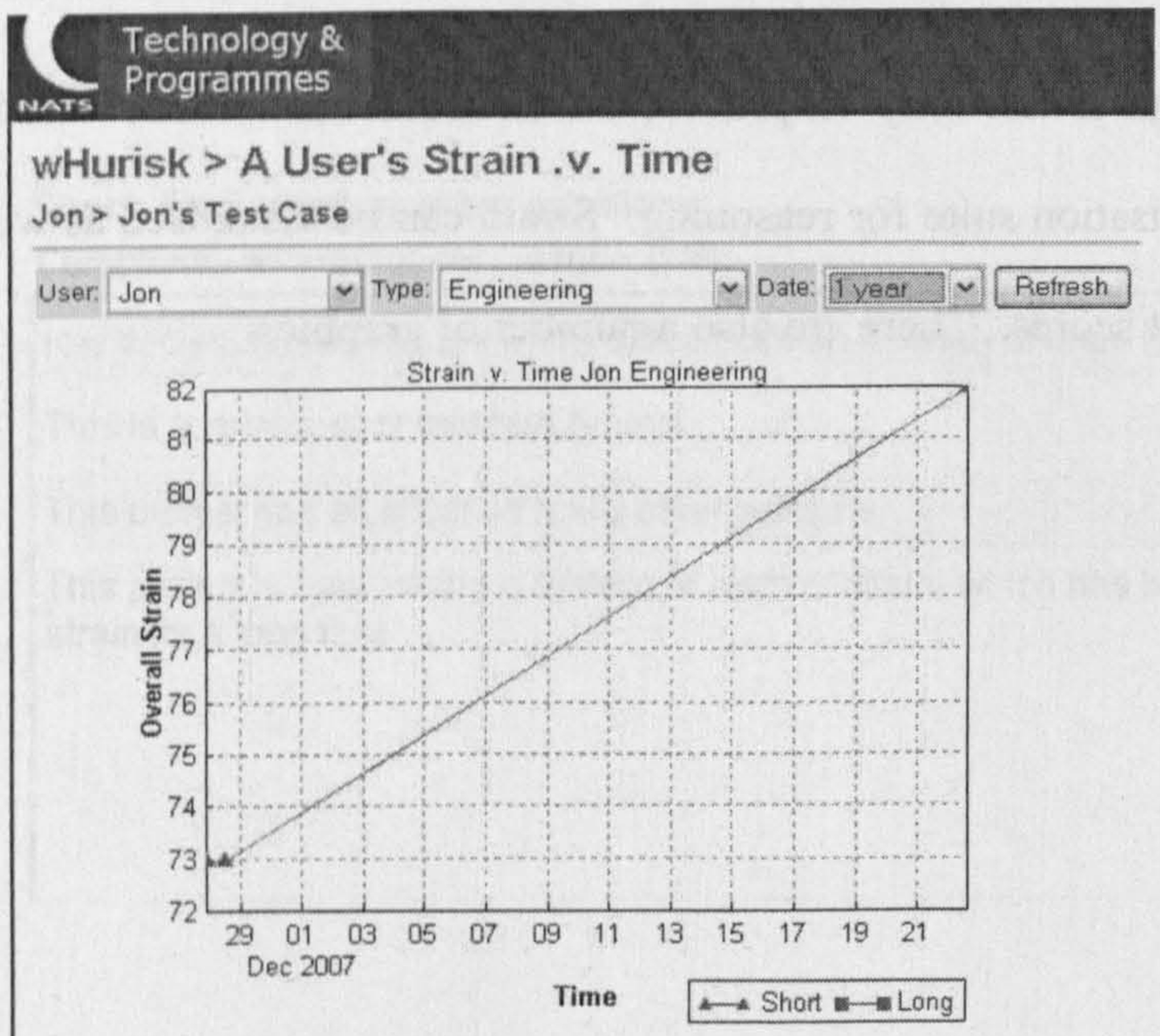
Figure 7: Strain visualisation options



The following examples show how the system helps users to home in on key ideas:

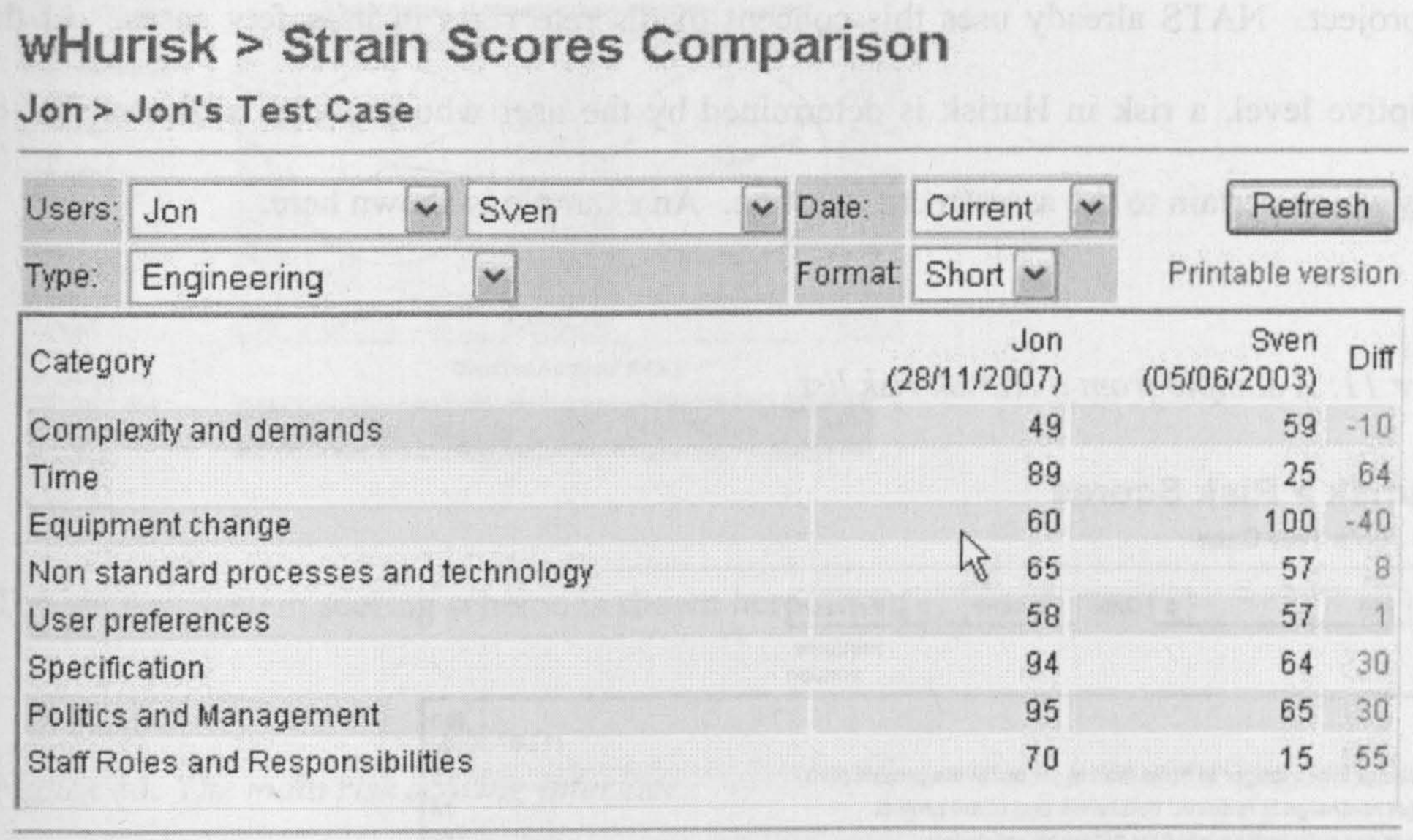
Idea one: where is the strain in this project going?

Figure 8: Comparing changes in strain over time (single user)



Idea two: What do I think the key contributors of strain in this project are and how does that compare with my colleagues views of the same?

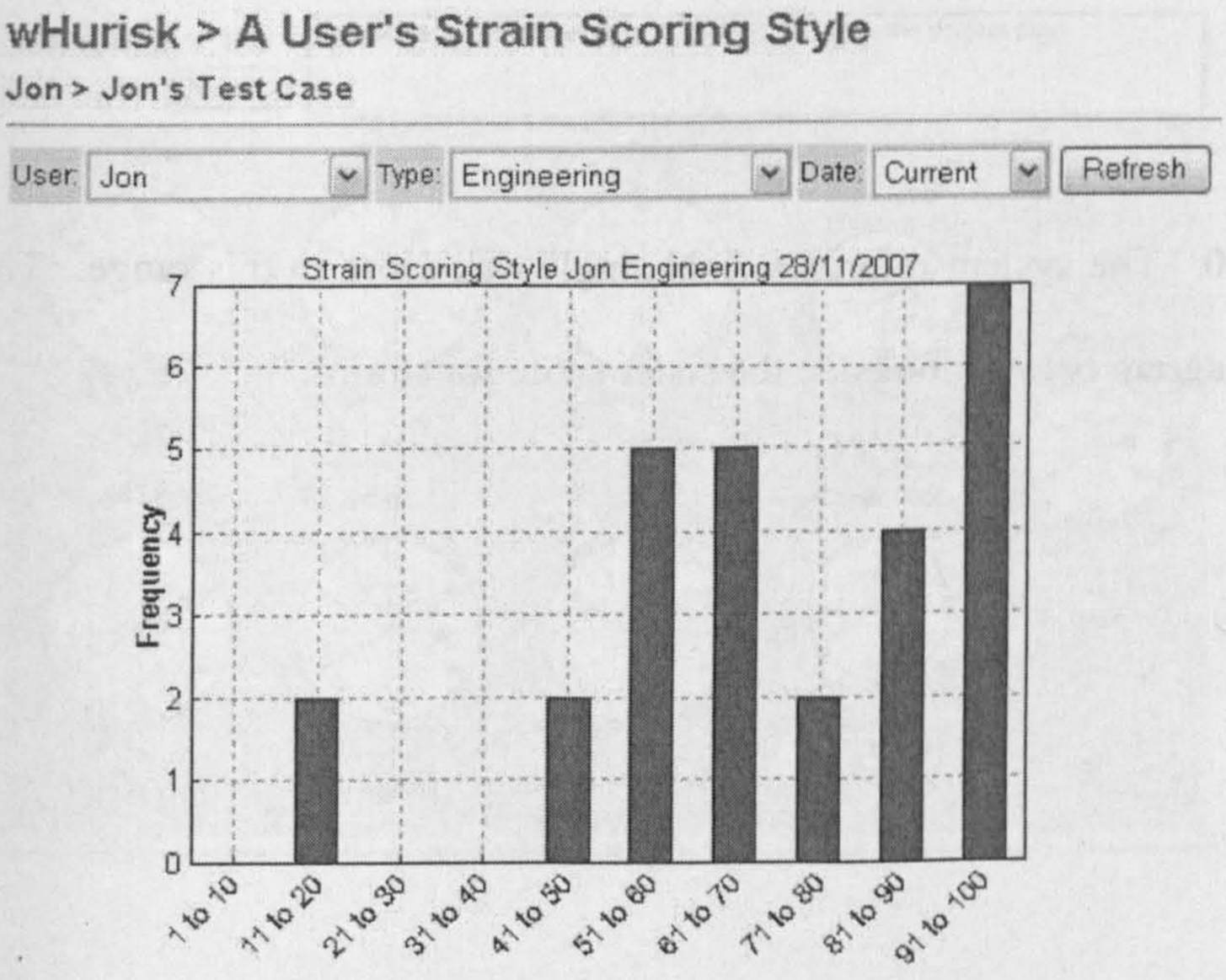
Figure 9: Comparing different users views on strain at the category level



Category	Jon (28/11/2007)	Sven (05/06/2003)	Diff
Complexity and demands	49	59	-10
Time	89	25	64
Equipment change	60	100	-40
Non standard processes and technology	65	57	8
User preferences	58	57	1
Specification	94	64	30
Politics and Management	95	65	30
Staff Roles and Responsibilities	70	15	55

Idea three: how do I generally score strains when I do?

Figure 10: Assessing scoring style



Risk measurement and visualisation

Risks are simplified for the purposes of Hurisk. Simply put, they are a summarisation of negative events which may occur to upset safe operational functioning or the safe execution of a project. NATS already uses this concept of discrete risks in its safety cases. At the descriptive level, a risk in Hurisk is determined by the user who provides a discrete set of events which pertain to the assessment question. An example is shown here.

Figure 11: A sample from a Hurisk risk list

wHurisk > Risk Scores

jon > Jon's Test Case

User: Jon Date: Current Refresh

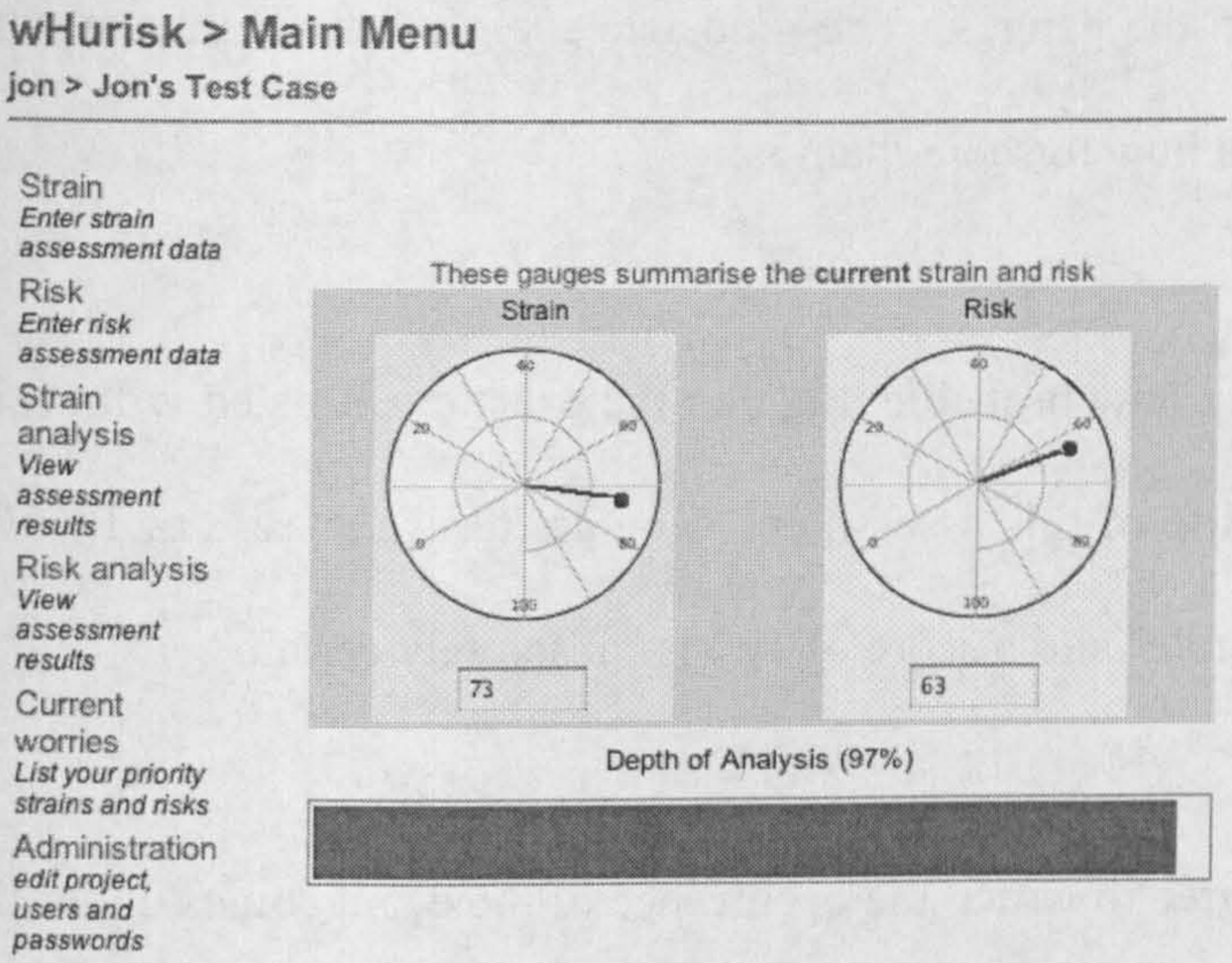
Printable version

Risk	jon (13/07/2005)
The trainer takes longer to build than is set out in the project plan	75
A major re-design is required before the end of the project	74
There is not sufficient time left for thorough validation	50
The trainer does not overcome the drawbacks of the Hillway Knee	75
The Hull simulated is more advanced / sucessful	50
The level of registration between physical and virtual components of the design does not have a high enough fidelity	75
The tissue deformation is not realistic	75
"VR research does not answer the how real is real qestion	51

Inputting risk measures

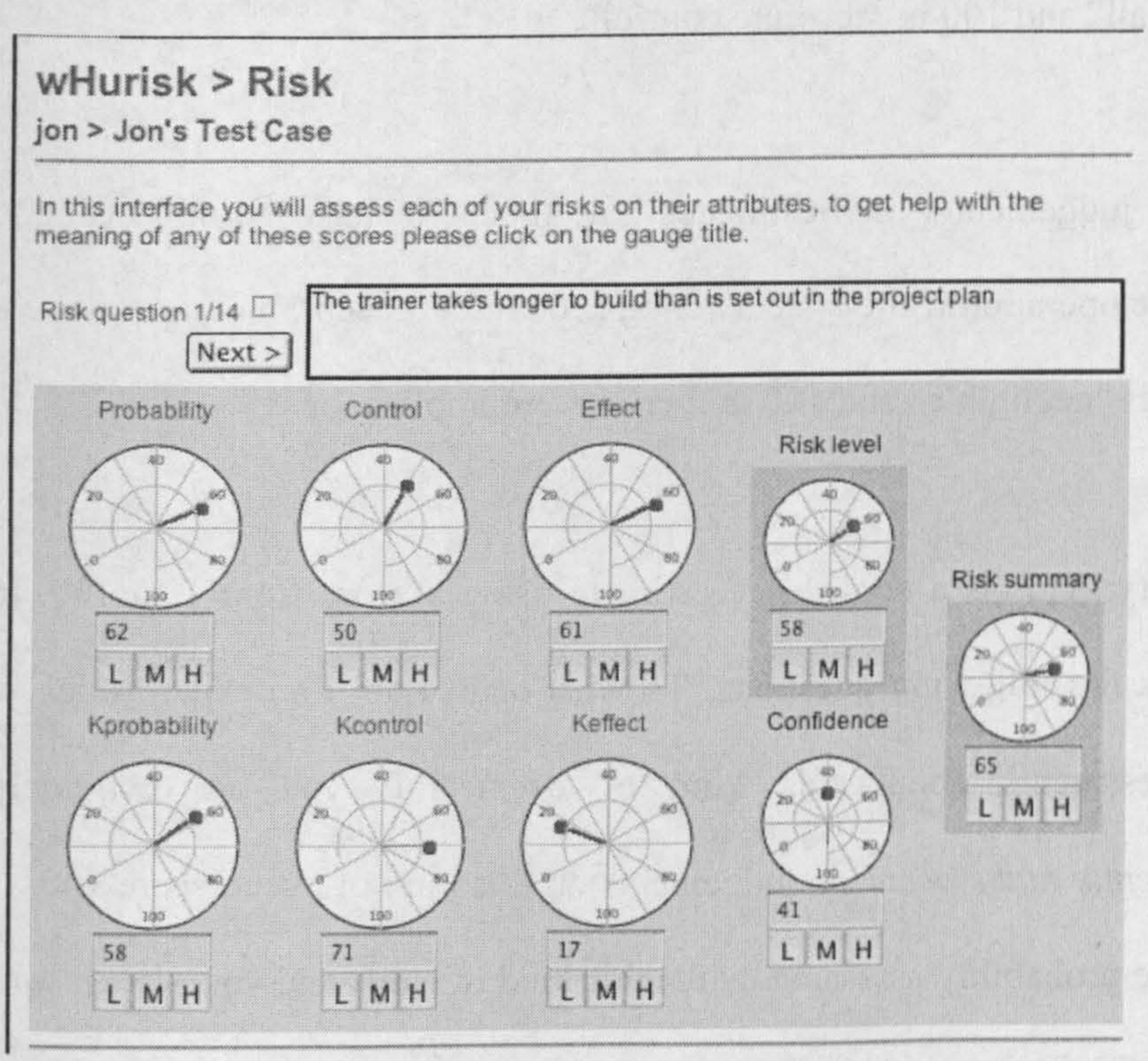
Hurisk scores are all 1-100. The system returns a final single risk score in this range. That score can be seen in the diagram below alongside the sister score for strains.

Figure 12: Summary page of the system



The six part system scoring scheme is shown here

Figure 13: The main risk scoring interface



Hurisk expects the user to supply a risk score as 'L,M, or H' as a discrete score including direct manipulation of the needles on the gauges. The guidance given by the Hurisk system on the meaning of these elements and how to score them is:

Probability: This score is estimating how probable it is that the exact event listed will occur within the life of the project or the life of this assessment. You are to use a scale of 1 - 100 where a score of 1 is "almost impossible" and a score of 100 is "absolutely certain".

Control: This score refers to the degree to which the occurrence of the event comes under the control of you as a single agent or your group in the organisation (e.g. your ability to control whether or not you contract an 'immunable' disease would be rated against your access to immunisation). You are to use a score of 1 to 100 where 1 indicates "no control over the occurrence of the event at all" and 100 is "absolute control".

Effect: This is a severity judgement which estimates how much "damage" would occur to your project or your future operation if the exact event listed were to occur. You are to use a score of 1 to 100 where 1 is "negligible" and 100 is "very severe implications".

Knowledge of probability: This is a confidence estimate using a score from 1 to 100, to score your confidence ask the following question: "To what degree am I an expert judge in making probability assessments about this kind of event?" If you are extremely knowledgeable and *frequently make accurate* judgements on this kind of event, score 100. If you do not normally make probability assessments for this kind of event and you are *not sure how accurate you are* in so doing, score 1.

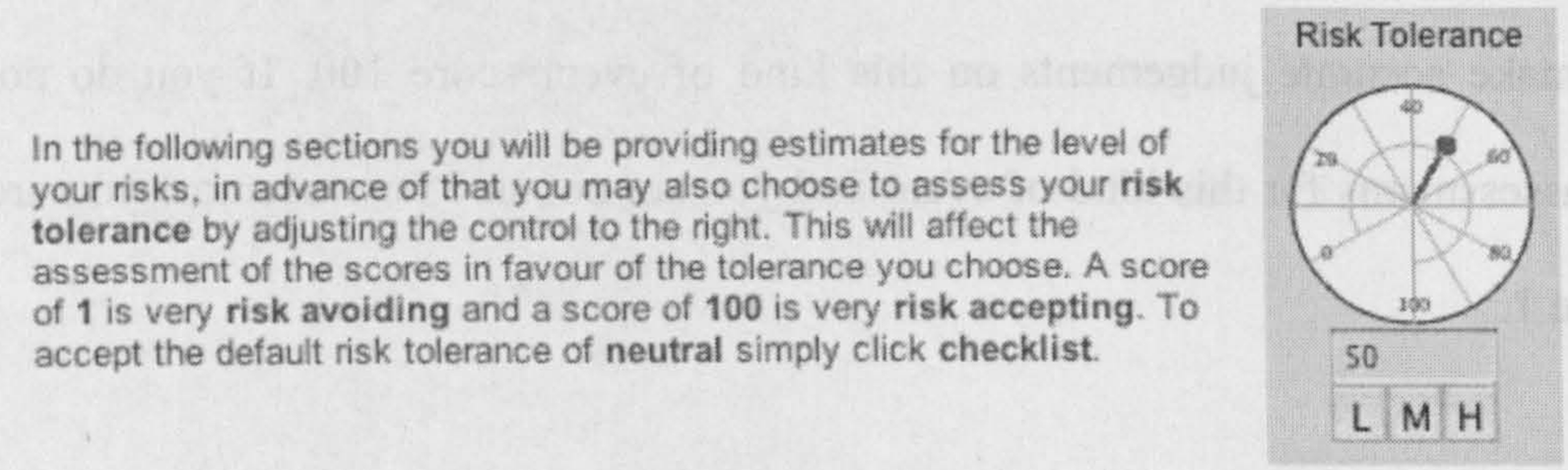
Knowledge of 'controlability': This is a confidence estimate using a score from 1 to 100, to score your confidence ask the following question: "To what degree am I an expert judge in making assessments of control over this kind of event?" If you are extremely knowledgeable and frequently make accurate judgements on this kind of event score 100. If you do not normally make assessments for this kind of event and you are not sure how accurate you are in so doing, score 1.

Knowledge of effect: This is a confidence estimate using a score from 1 to 100, to score your confidence ask the following question: "To what degree am I an expert judge in making assessments about the effect of this kind of event?" If you are extremely knowledgeable and frequently make accurate judgements on this kind of event score 100. If you do not normally make effect assessments for this kind of event and you are not sure how accurate you are in so doing, score 1.

Risk tolerance

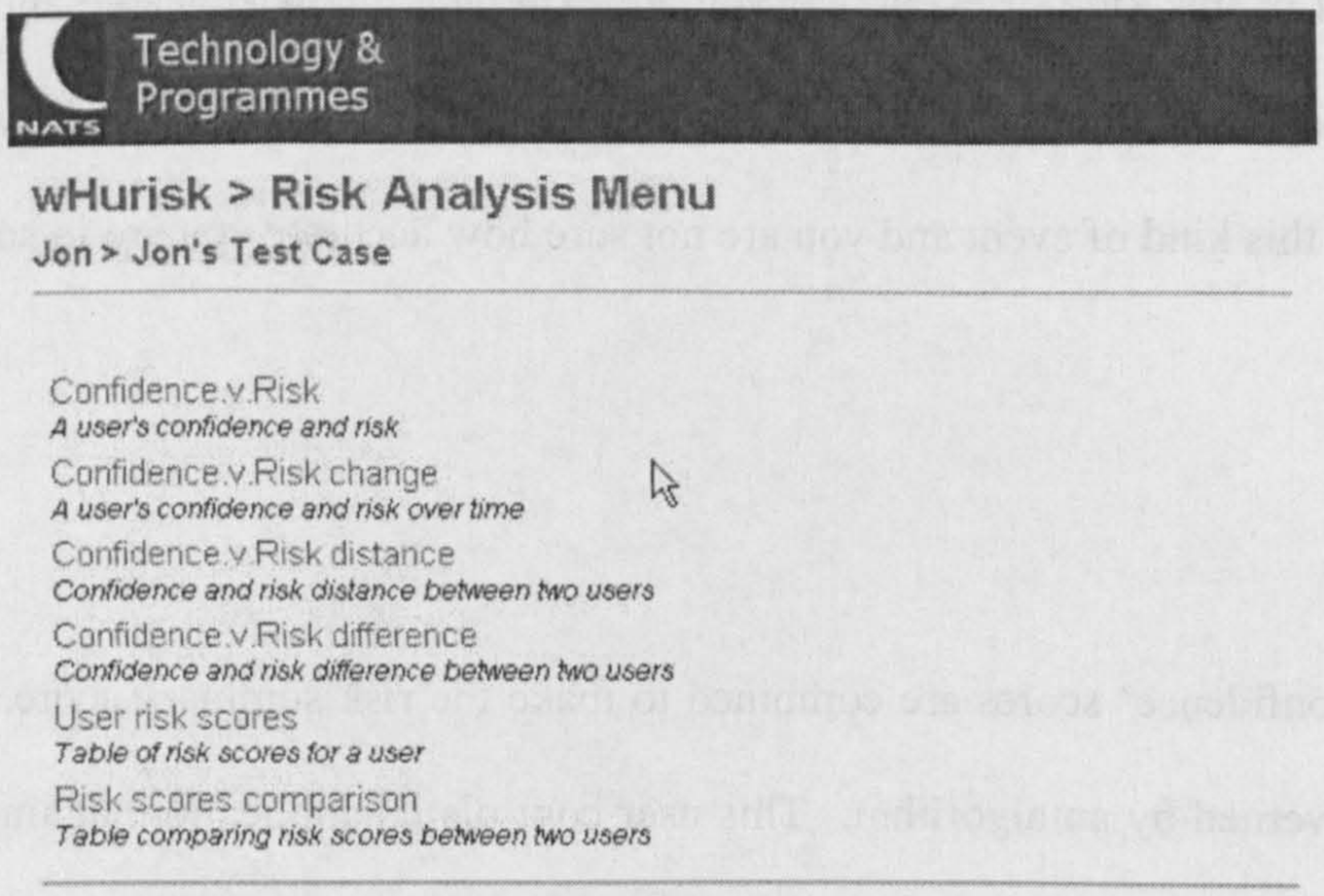
The 'Risk level' and the 'Confidence' scores are combined to make the risk summary score. That combination rule is governed by an algorithm. This user controls its application, again by applying a score of 1-100 for a variable elsewhere in the system called "risk tolerance".

Figure 14: Risk tolerance gauge and instruction



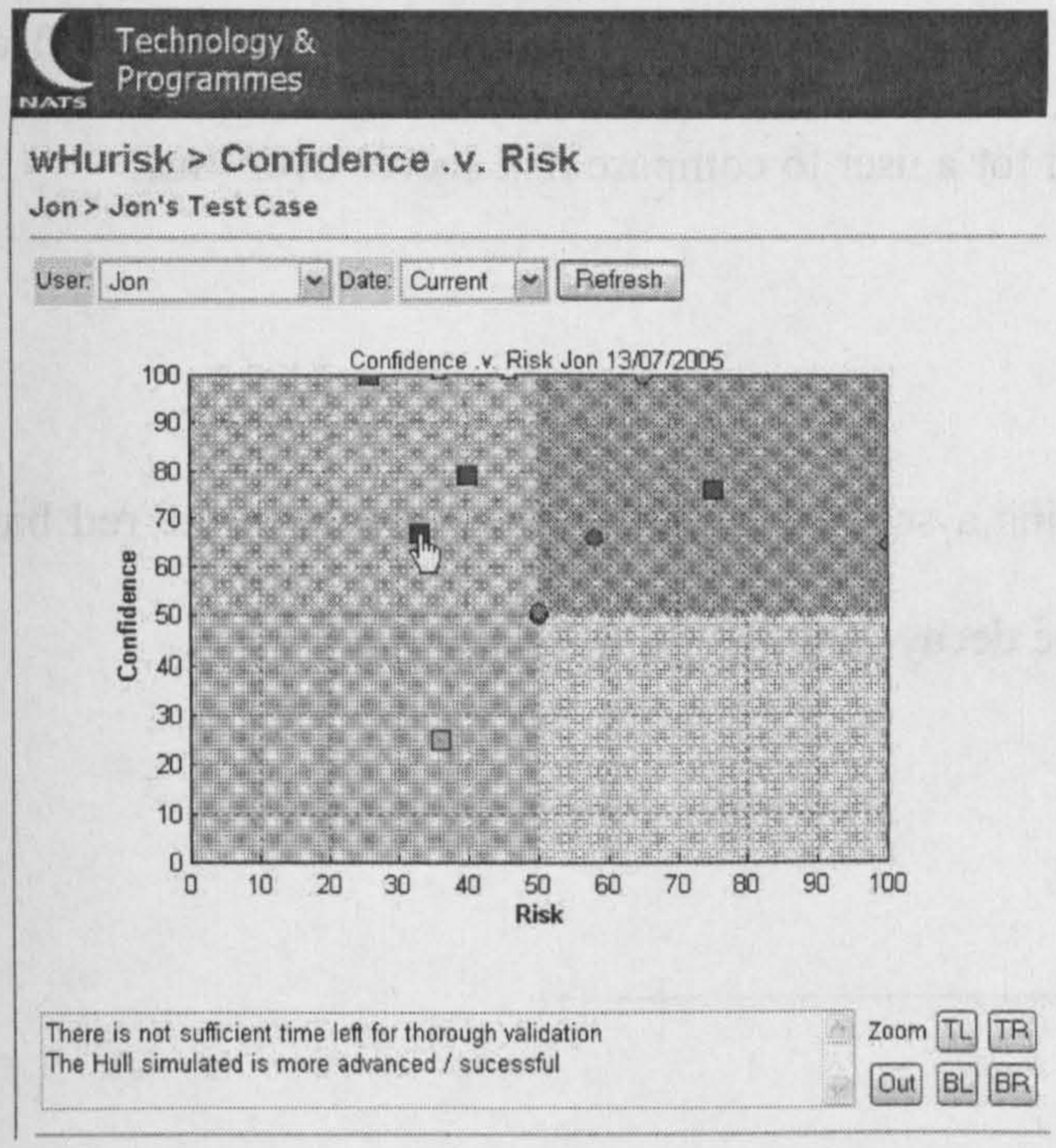
Visualising risk

Figure 15: Visualisation options for risk



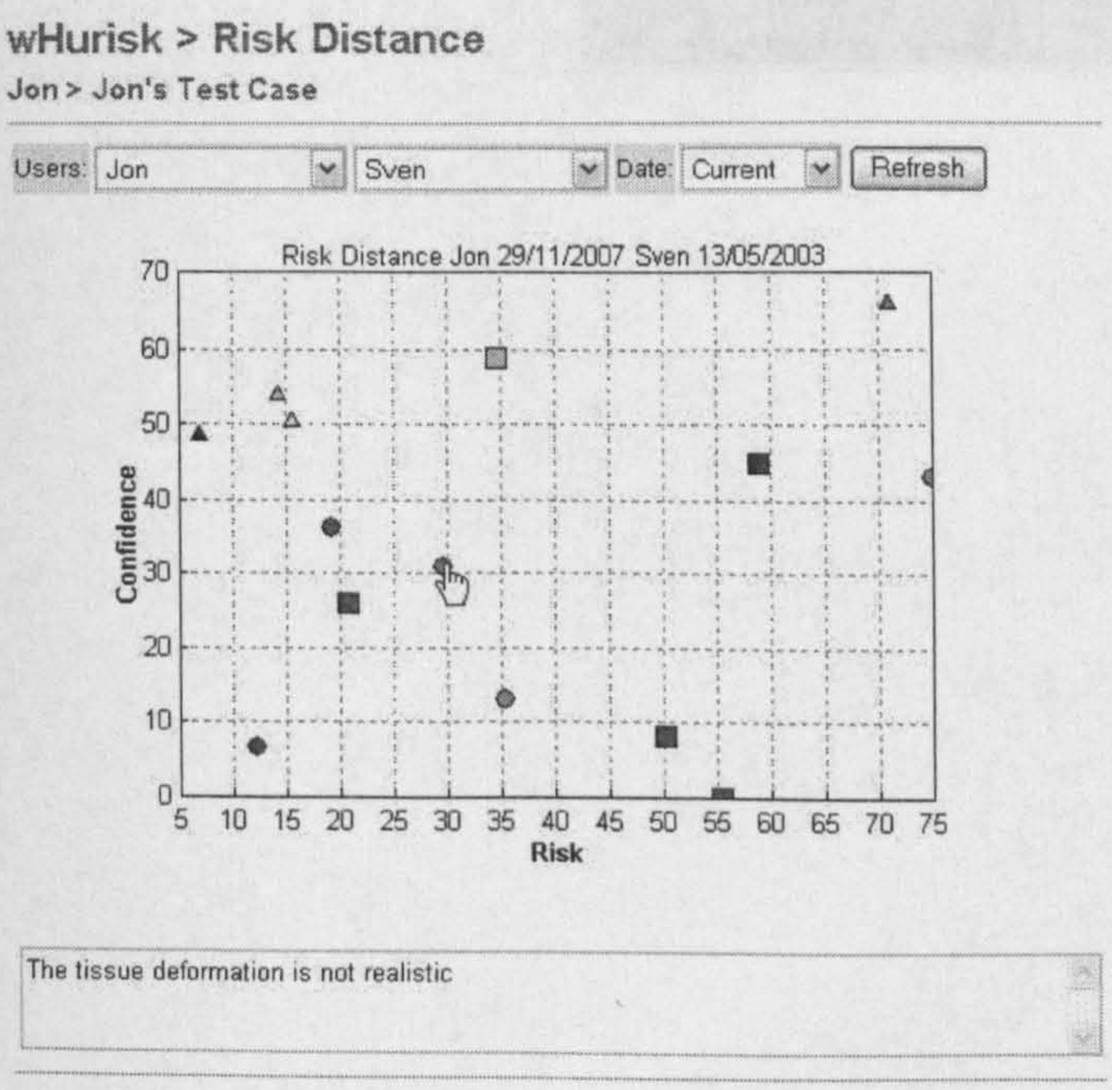
Hurisk allows a number of visualisations of the risk data. Two examples will be shown for illustration. First, confidence plotted against risk summary. As can be seen here the user explores the relative positions of the risks in a two dimensional plot either just for the current data, as shown, or by selecting two time periods for comparison.

Figure 16: Risk and confidence



One of the functions is to look at the distance between two users' risk scores when they have performed a separate analysis of the same set of risks. This visualisation helps you see how discrepant the scores are and over what. This is one of the key decision aids focussing attention to the right things.

Figure 17: Risk distance plotting

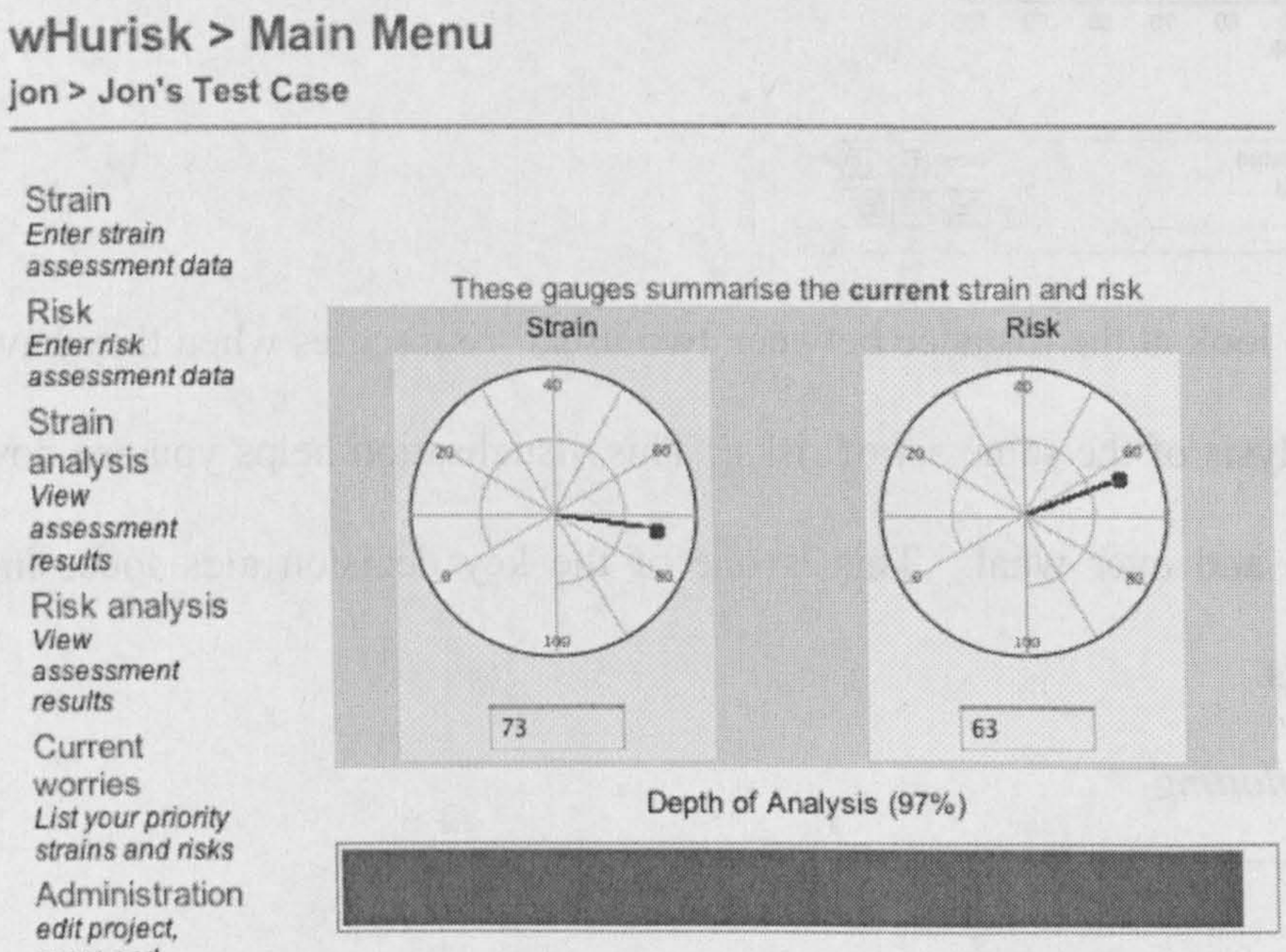


The calculation do this a simple two dimensional Euclidian distance calculated on the confidence and risk parameters and plotted in two dimensions. This Euclidian distance is also used when the same visualisation is utilised for a user to compare risk scores over time.

The audit bar (decay function)

As part of the visualisations Hurisk contains a schoolteacher function shown in the red bar here. This depth of analysis is subject to the decay function described earlier.

Figure 18: Highlighting depth of analysis



3.7.4 Lesson learning measurement ideas and visualisation

Introduction

Hurisk had a number of lesson learning prototypes within it. This remained, at the end of the project, the least developed area and the least easy one to provide any tool concepts for. For completeness, three approaches are interesting to consider for the purposes of this thesis. It is accepted at the outset that these are too soft to draw very effective conclusions from. In consequence of this the measurement and visualisation ideas will be considered in one combined section.

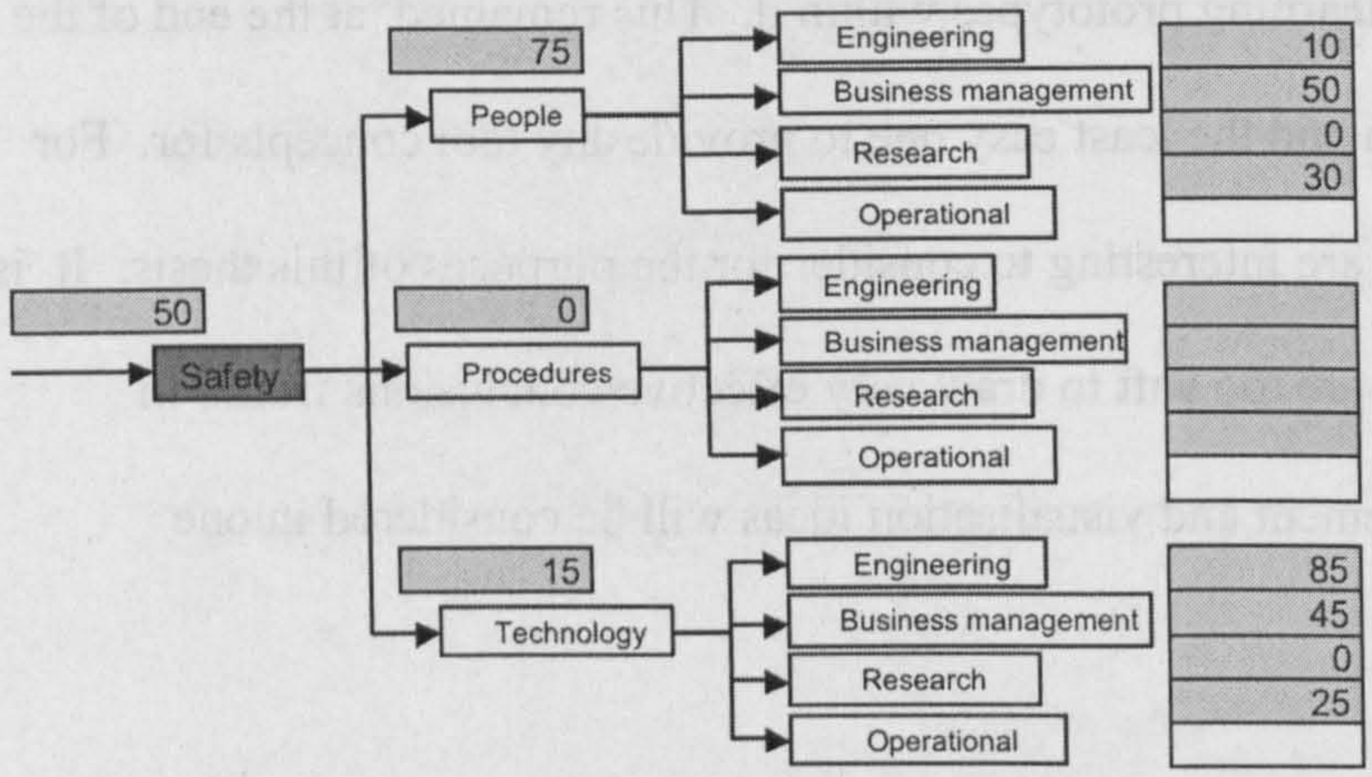
The techniques covered comprise of a fresh version of the triads from Hurisk one, a set of matrix-based questionnaires with visualisations and a function called "worry beads".

Taking the triads out again

In a development of the triads from Hurisk 1 (Safety, Efficiency, Capacity & People, Procedures, Technology) a third triad called "focus" was added (Engineering research focus, business focus, operations focus). A glorified spreadsheet was used to collect data on specific projects. The safety arm is shown below. A group of decision makers would agree on the "parity" of the object to be analysed. This ranged from a single decision to an entire project. To provide scores users allocate 1-100 for the influence they feel these factors have.

Scoring and presenting the triad diagram

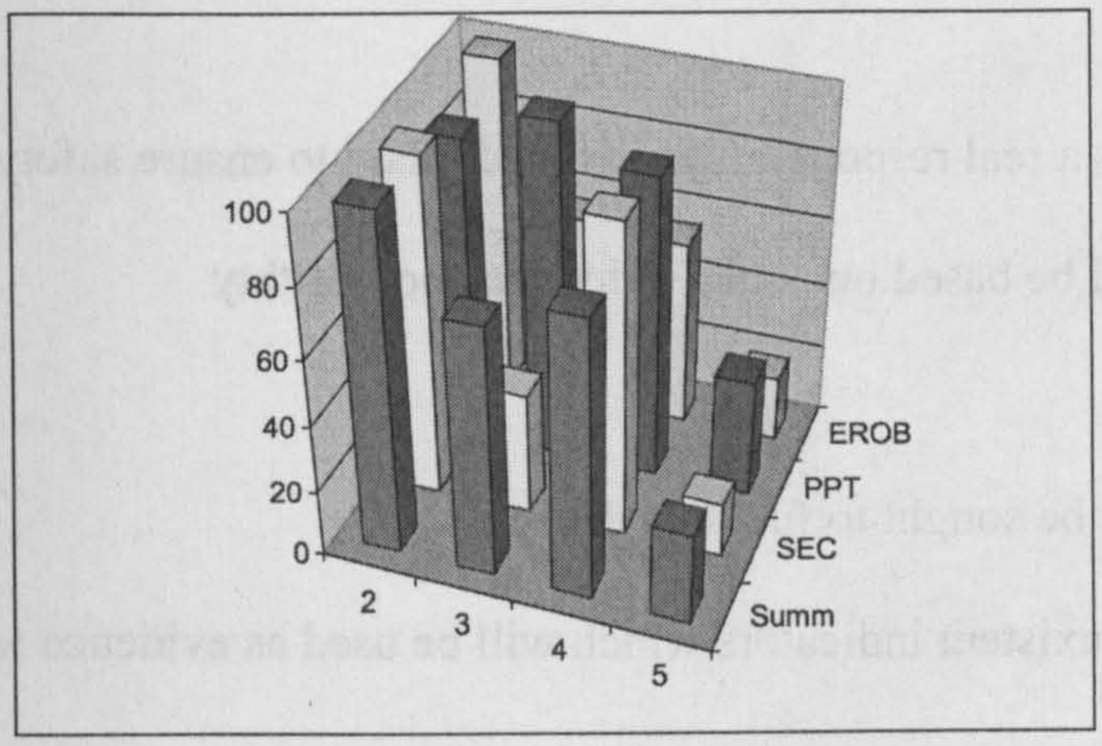
Figure 19: The triad scoring diagram



This diagram has three levels, but the data treatment is the same for all levels. Using these scores the three dimensional Euclidian distance between decision makers is calculated at each level. In the case of a summary score to represent distance at all levels the nine dimensional Euclidian distance is used. These distances are then re-scaled. The largest distance is set to 100 to indicate the largest proportional disagreement and the others are scaled against that.

Data is presented back to each individual decision maker in a "city plot" format, an example is shown below. The decision maker reading the plot is always at the zero point. From this plot the decision maker can quickly understand who disagrees with them the most and over what. The triad diagram is also available of course for any "deep dive" activity.

Figure 20: Decision distance city plot



Matrix based questionnaires

Two tools were used experimentally with members of the NATS steering group to further develop lesson learning (or decision comparison) methods. The first tool was a highly conceptual safety taxonomy. It's core idea was that, particularly in the changing political situation of NATS at that time, the identity of the decision makers on projects had a bearing on the safety-consciousness of the decision making. A number of criteria were developed partly from the Hurisk one risk analysis and partly from the experiences of the steering committee. These criteria were scored again in a simple spreadsheet environment which looked like this.

Figure 21: A safety consciousness taxonomy

Categories		PW	MR	JS	PW - JS	MR - JS
CORE REALITY	Concrete	100	50	85	15	-35
	Abstract	1	50	40	-39	10
	Real responsibility for safey	100	2	50	50	-48
	Direct physical measures (*of safety*)	75	60	98	-23	-38
MEASURES AND STIPULATIONS	Data processing used	10	1	100	-90	-99
	Indicator based	10	50	100	-90	-50
	Procedurally led	100	100	100	0	0
	Engineering biased	10	10	10	0	0
PERSONAL ENGAGEMENT	Experiential element	50	1	1	49	0
	Subjective qualification	100	50	50	50	0
	Emotional element	1	75	100	-99	-25
COMPREHENSION AND INFLUENCE	Complex responsibility	90	20	90	0	-70
	Delegated responsibility	100	1	1	99	0
	Flexible definition	90	20	60	30	-40
	Market led	100	1	50	50	-49

Core reality:

- The degree to which the objects in this decision are real, tangible and fixed

- The degree to which the elements of this decision are intangible, abstract and prone to change meaning over time
- The degree to which this agent has a real responsibility in this domain to ensure safety
- The degree to which decisions will be based on actual evidence about safety

Measures and stipulations:

- The degree to which evidence will be sought to fuel this decision
- The degree to which there are pre-existent indicators which will be used as evidence to fuel this decision
- The degree to which this decision will be led by what is feasible to do within a standard operating procedure
- The degree to which the measurements which are being utilised to aid this decision are sourced in the engineering world

Personal engagement:

- The degree to which the decision maker will have real experiences of the aspects of this decision both now and once it is made
- The degree to which the decision maker is likley to rely on their own subjective judgements about this decision
- The degree to which the decision maker is emotionally tied up with the subject matter of the decision

Comprehension and influence:

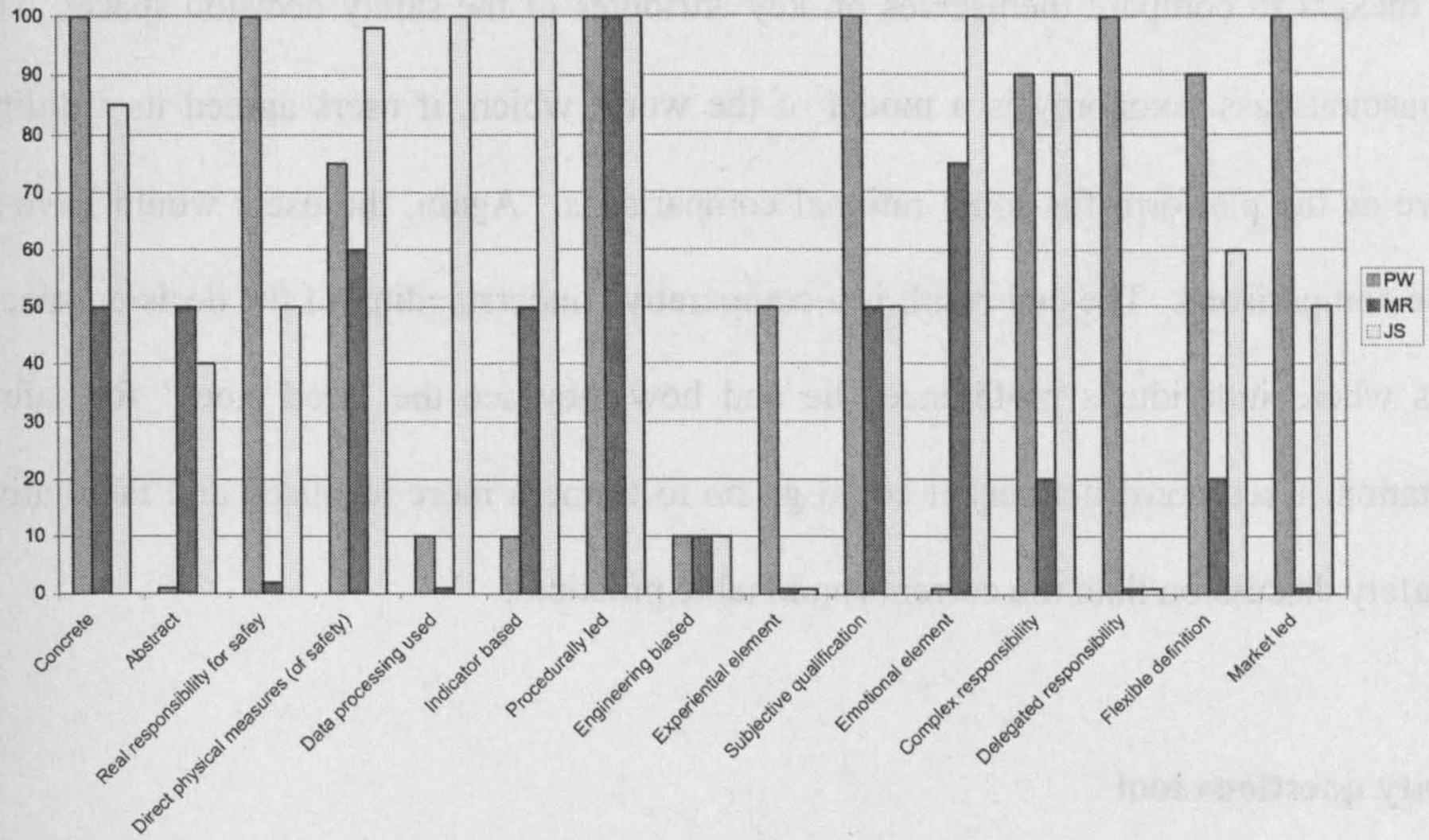
- The degree to which the decision maker understands the world of this decision from a position of having complex responsibilities in it
- The degree to which the decision maker, once this decision is made, delegates its effects to another agent
- The degree to which this decision maker has to be flexible in their understanding of safety in order to make judgements and compromises about it

- The degree to which this decision maker will be influenced by market, profit and customer satisfaction issues in making this decision

The user first scores themselves and then at least one other decision maker. The decision maker scores are subtracted from the user scores and the results are plotted in a 'tornado style' bar chart where the user can see disagreements in terms of magnitude and direction.

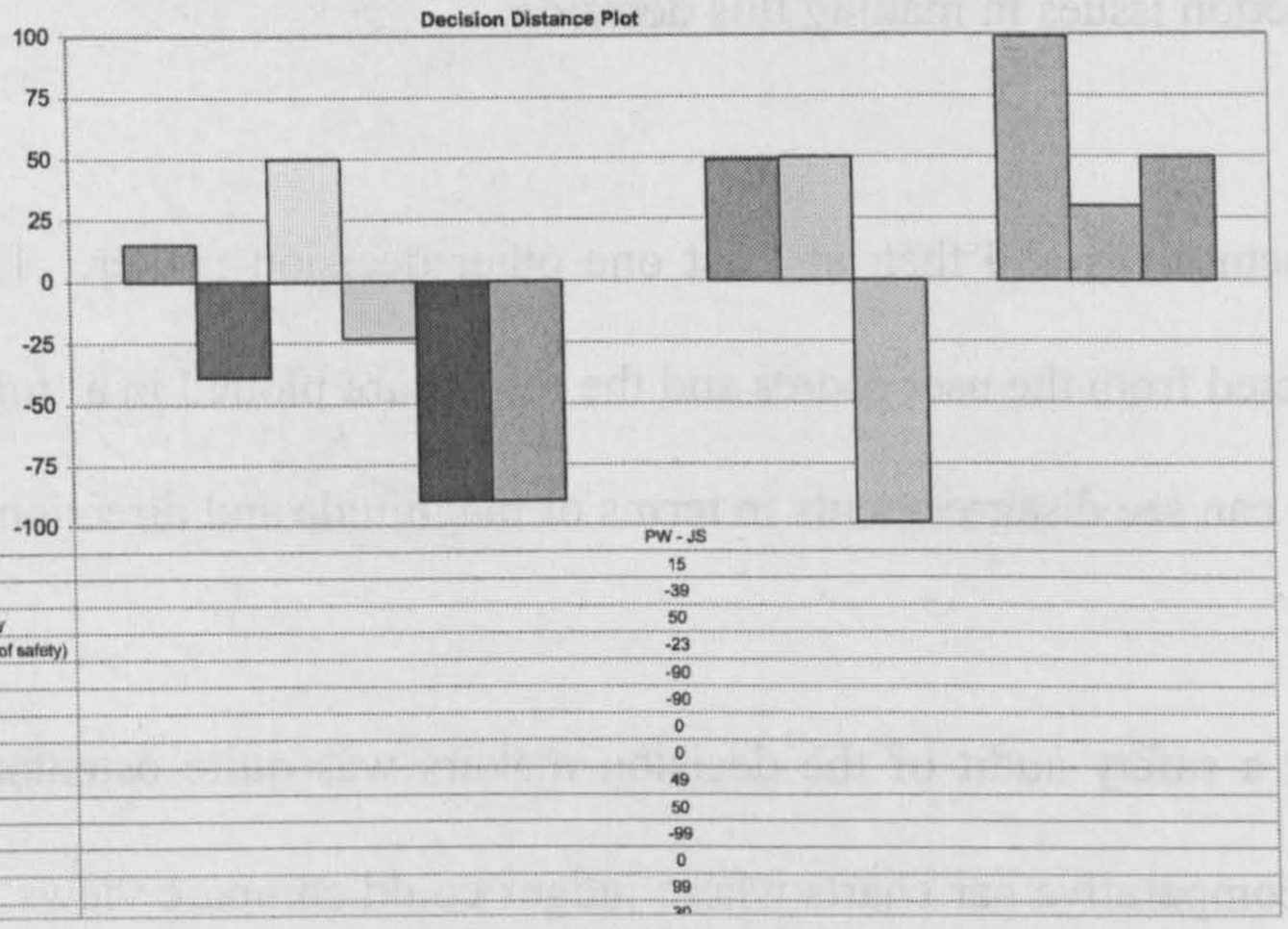
The idea of performing a safety audit of the decision makers was quite ostentatious. The results were a series of comparative bar charts where judges could compare views. Examples are shown:

Figure 22: safety taxonomy bar chart



Disagreement as distance between two users was also plotted

Figure 23: The decision distance plot



	15
Abstract	-39
Real responsibility for safey	50
Direct physical measures (of safety)	-23
Data processing used	-90
Indicator based	-90
Procedurally led	0
Engineering biased	0
Experiential element	49
Subjective qualification	50
Emotional element	-99
Complex responsibility	0
Delegated responsibility	99

Taking stock of this approach

This approach is aimed at decision support, to give a clearer and more structured way for decision makers to compare themselves on key attributes of the safety decision space. The safety consciousness taxonomy is a model of the world which, if users agreed its validity, does serve as the platform for some rational comparisons. Again, the users would have to accept the comparisons. The end result is a comparative understanding of the decision space. It reveals where individuals preferences lie and how they see the 'seed stock' for safety argumentation. Used constructively it could go on to frame a more idealised and facilitative kind of safety discussion than the current round table practices.

The twenty questions tool

Although considered worthwhile the safety taxonomy had a number of drawbacks in terms of its comprehensibility and its highly subjective nature. The question was posed, why not stop being so fancy and just ask the blunt questions which the taxonomy is alluding to? The result was the 'twenty questions' tool (the number was a pure accident). Naturally, when you take this approach, what you end up with is a questionnaire which in itself is not very interesting from the point of view of developing decision support. But it was designed and considered.

This questionnaire elicits 20 scores from 1-100 for each decision maker which are arranged into five category groups of four questions. The analysis allows comparison between decision makers at the individual question or category level i.e. there are either twenty differences or four.

As a primer tool this questionnaire could set the landscape for decision making within a more demonstrable common ground which is more disciplined and compartmentalised. Users found the questionnaire to not only be highly provocative, but highly intuitive at highlighting issues of "focus", which key decision makers or key functional groups might have.

Figure 24: The twenty questions tool

TWENTY QUESTIONS. . .			
	Decision Maker :		
Understands	This decision maker fully understands the decision in terms of both its concrete and abstract elements		
	This decision maker is being guided by the formal analysis of real data in aspects of the decision which need that approach		
	This decision maker is monitoring real and appropriate indicators to inform their understanding of this decision		
	This decision maker has valid experience of the area in which this decision will have its effect		
	This decision maker does not have a personal bias based on their own preference for the way this decision should go		
In the fog	This decision maker is likely to rely on and be guided by their own subjective judgements when contributing to this decision		
	This decision maker is emotionally tied up with the content, the lifecycle and the outcome of this decision		
	This decision maker understands the world of this decision from a position of having complex responsibilities in it		
	This decision maker, once this decision is made, delegates all of its effects to other people		
	This decision maker will be influenced by market, profit and customer satisfaction issues in making this decision		
In the clear	This decision maker has a real responsibility in this domain to ensure safety		
	The decision will be based on valid and recent evidence about safety without any spin		
	The decision will be based around indices of safety which are predominantly physical		
	This decision maker does not have to be flexible in their absolute understanding of safety in order to make judgements within a bigger picture		
	This decision maker will not be influenced by market, profit and customer satisfaction issues in making this decision		
In the purse	This decision will have a significant impact on this decision maker's, or their area's, budget		
	This decision would ideally need, but does not come with, sufficient additional resources		
	This decision maker is likely to want to see additional safety justified in terms of a cost benefit analysis		
	This decision maker is likely to want to say a 'no' decision still conforms to 'expected standards' or compares favourably to *operations elsewhere*		
	This decision maker is required to or wants to save money		

Re-visiting the taxonomy

The final chapter in this part of the work was the decision to re-visit the safety taxonomy and attempt to strengthen it as a comparative profiling tool. The scoring was simplified to "yes / no" opinion and a matrix of "safety influence houses" were added:

- Operational practice: Safety was in the delivery of air traffic control
- Management practice: Safety in the management of the delivery of air traffic control
- Technology closed: Safety in "black box" technologies e.g. radar
- Business development: Safety in the face of remaining competitive and developing UK air transport
- The word: Safety enshrined in mandatory fixed written procedures e.g. MATS pt 11
- Tools: Safety mediated by interacting tools such as Terminal Collision Avoidance System

– Agents: Safety mediated by people

Some results from a trial of this approach are shown in the figure below. What one is looking for is high consensus areas which can quickly draw attention to the safety profile of a project or decision. In the example below the decision is notable because of the high contrast value between the 'operational'/ 'emotional' and 'business development' elements. Performing this analysis gave the users considerable food for thought in terms of how the safety bearing properties of these decisions might be driven.

Figure 25: an expanded safety considerations taxonomy

Categories		Elements & *sub elements*						
		Operational practice	*Management practice*	*Technology closed*	*Business development*	The Word	Tools	Agents
CORE REALITY	Concrete	Yes	Yes	Yes	No	Yes	Yes	Yes
	Abstract	No	Yes	No	Yes	Yes	No	Yes
	Real responsibility	Yes	Yes	No	No	No	No	Yes
	Direct physical measures (*of safety*)	Yes	No	No	No	No	Yes	Yes
MEASURES AND STIPULATIONS	Data processing used	Possible	Yes	Yes	Yes	No	Yes	Yes
	Indicator based	No	Yes	No	Yes	Yes	Yes	Yes
	Procedurally led	Yes	No	No	No	Yes	Yes	No
	Engineering biased	No	Possible	Yes	No	Yes	Yes	Possible
PERSONAL ENGAGEMENT	Experiential element	Yes	No	No	No	No	No	Yes
	Subjective qualification	Yes	Yes	Possible	Yes	Yes	Possible	Yes
	Emotional element	Yes	No	No	No	No	Yes	Yes
COMPREHENSION AND INFLUENCE	Complex responsibility	No	Yes	No	Yes	No	No	Yes
	Delegated responsibility	No	Yes	No	Yes	Yes	Yes	Yes
	Flexible definition	No	Yes	No	Yes	Possible	Yes	Possible
	Market led	No	Possible	Possible	Yes	No	Yes	No

How this approach could work is as an early warning system for arguing at cross purposes giving as it does an early landscape of the whole decision space. The use of such a taxonomy could conceivably cause discussions that might otherwise not have taken place (or have the permission to be). In this way, like all of the lesson tools, it's commendable to set a frame for dialogue but it is hard to see it as "a tool" in the sense which I have come to discuss them.

Worry beads

The Hurisk concept demonstrator contained the well developed risk and strain models highlighted earlier. As discussed above the lesson learning / decision modules were not maturing quickly enough out of concept. There was much background modelling not shown here and a great deal of discussion about how tools could be built from these and other analytical ideas. What became clear though was that these would not be built.

A desire to have some sort of lesson learning built into the tool for decision support remained however. Reviewing what the system could already do led us to the idea that collecting the priority risks and strains from any given project might serve as a good proxy for lesson learning if they could be carried forward to other similar projects. Functionality was added to Hurisk to export these.

Given the length of many NATS projects, it was then decided that the lessons one could learn from the high priority risks and strains might actually add intuition within the life of the project as well as retrospectively. Reviewing top of table risks and strains was considered a valuable activity. The worry beads make this recognised good practice easy to achieve.

To facilitate this the 'worry beads' viewer was introduced. The user set the filtering threshold (strains or risks above 75, 85, 90 %). The figure below illustrates how this looked.

Figure 26: Worry beads

wHurisk > Current Worries

jon > Jon's Test Case

This screen lists your current worries, these are a sub set of your 'strain' and 'risk' data covering the top 10 -25 % of your scores. These are thus the things which worry you at present

Strain threshold: 90 Refresh Printable version

The technology in this project may throw up issues which will be very time consuming	100
There are complicated politics which have not been resolved	100
The later the re-specification the more it will be about trying to keep everybody happy	99
The specification process here is not realistic on a number of fronts	99
Re-specification is inevitable on a number of fronts	99
Re-specification will mainly be about money not user needs	99
The suppliers will take advantage of us if we have late re-specification	98

Risk threshold: 90 Refresh

the time factor in which is relatively unknown and could be very	90
The project becomes technical over training focussed	100
The final system does not provide proven effective training	100
Performance measurement in the system is not valid	100
Tutors don't accept the system	100
The system is not perceived by users as more beneficial than current training arrangements	100

Functionality which was mooted (but alas not built) was to make these worries into a rolling screen-saver to help users stay highly conscious of them.

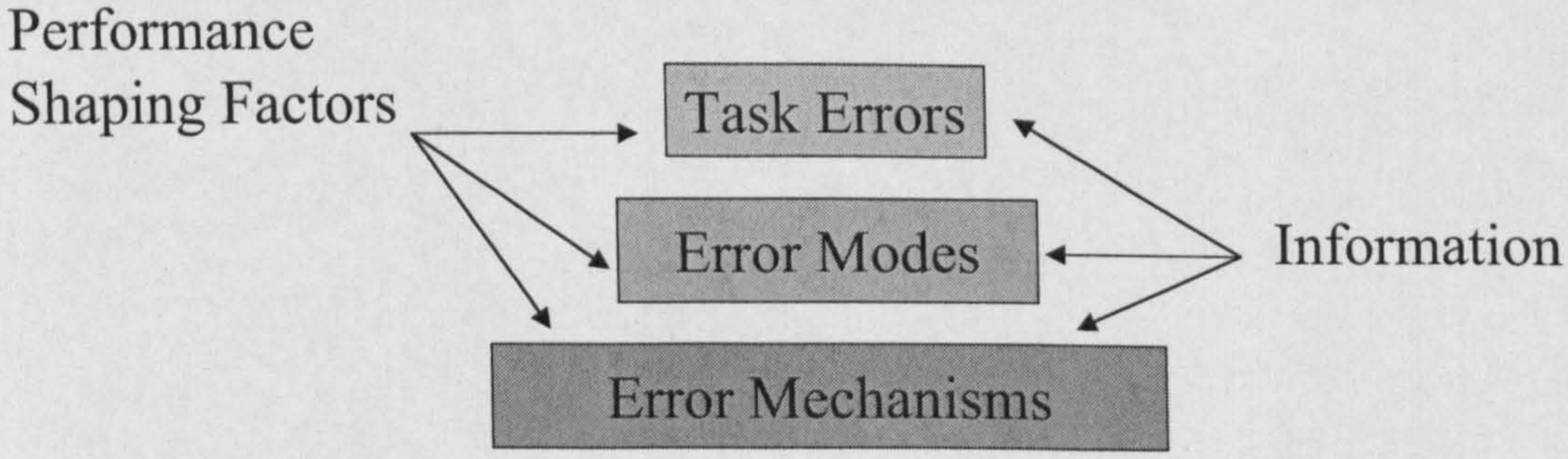
3.7.5 Benchmarking within NATS

Working with the Human Factors community in NATS, who took a keen interest in Hurisk, we decided to perform a benchmarking exercise looking at the closest thing they had yet developed to a heuristic system for reasoning support. Theirs was called Tracer and it acted as a guided reasoning tool to assist operational accident investigation. Tracer addressed the shift in root cause of air traffic incidents from system failures to human error (a generalised trend in air traffic) at the time in excess of 90% of incidents were caused by human failures - more specifically, human error.

The well developed processes for analysing failure modes within engineered systems were very easy for NATS deterministic systems being compliant with this form of enquiry. Recognising that human operators are not deterministic systems, a new way to mitigate failures was needed if a human error analysis was to work.

The basis for TRACEr was a decomposition of errors to assess which cognitive processes failed. Mitigation could then either reduce the number of future occurrences of the error, or reduce it's consequences the next time it happened.

Figure 27: TRACEr Hierarchy



The raw data for Tracer are in the form of incident reports produced by the individual who made the error, which can be backed up by interviews. The output is purely subjective and in narrative form. Tracer did not have a formal scoring system. The reason for this is

institutional and complex. The formal scoring of incidents lies elsewhere within the structure of NATS and uses and approved format which has not at this time, taken on the work of Tracer in identifying error pathways.

Heuristics

Tracer and Hurisk can both be designated as heuristic tools. In the Tracer case this is because of the way in which the tool is to be used. In the Hurisk case this was a design issue from the start. Neither tool provides deterministic or directive data, such as 'answers'. Each tool provides the capacity to look at pertinent data and, importantly, to track it over time. In the Tracer case this data remains raw. The incident investigator uses qualitative data to develop a comprehensive view of the complex issues within the incident.

Tracer can also be considered heuristic because it is designed to comply with a process which is essentially one of expert judgement. The four stages of incident investigation outlined by the Health and Safety Executive (HSE) guidance (HSG65 – Successful Health and Safety Management) are:

- Collect evidence about what has happened
- Assemble and consider the evidence
- Compare findings with appropriate legal, industry and company standards
- Implementation of findings and tracking of progress

Current practice in this area is a professional judgement Tracer helped the investigator consider all relevant data. Tracer fits further into the heuristic mould because it considers a cocktail of active failures (e.g. missed actions) and latent failures (e.g. poor or incomplete training, bad design). These latter failures are inextricably mixed with the human experiences of the operator and the human environment of the operating control room. They are necessarily softer factors and do not present themselves easily, even through the use of fault

trees, hazard & operability analysis or Failure modes analysis and the like, to meaningful quantification. Also, Tracer is aimed at reducing or mitigating effects of errors which occur in a human technical system which is dynamic, sometimes indeterminate and contains unknowns.

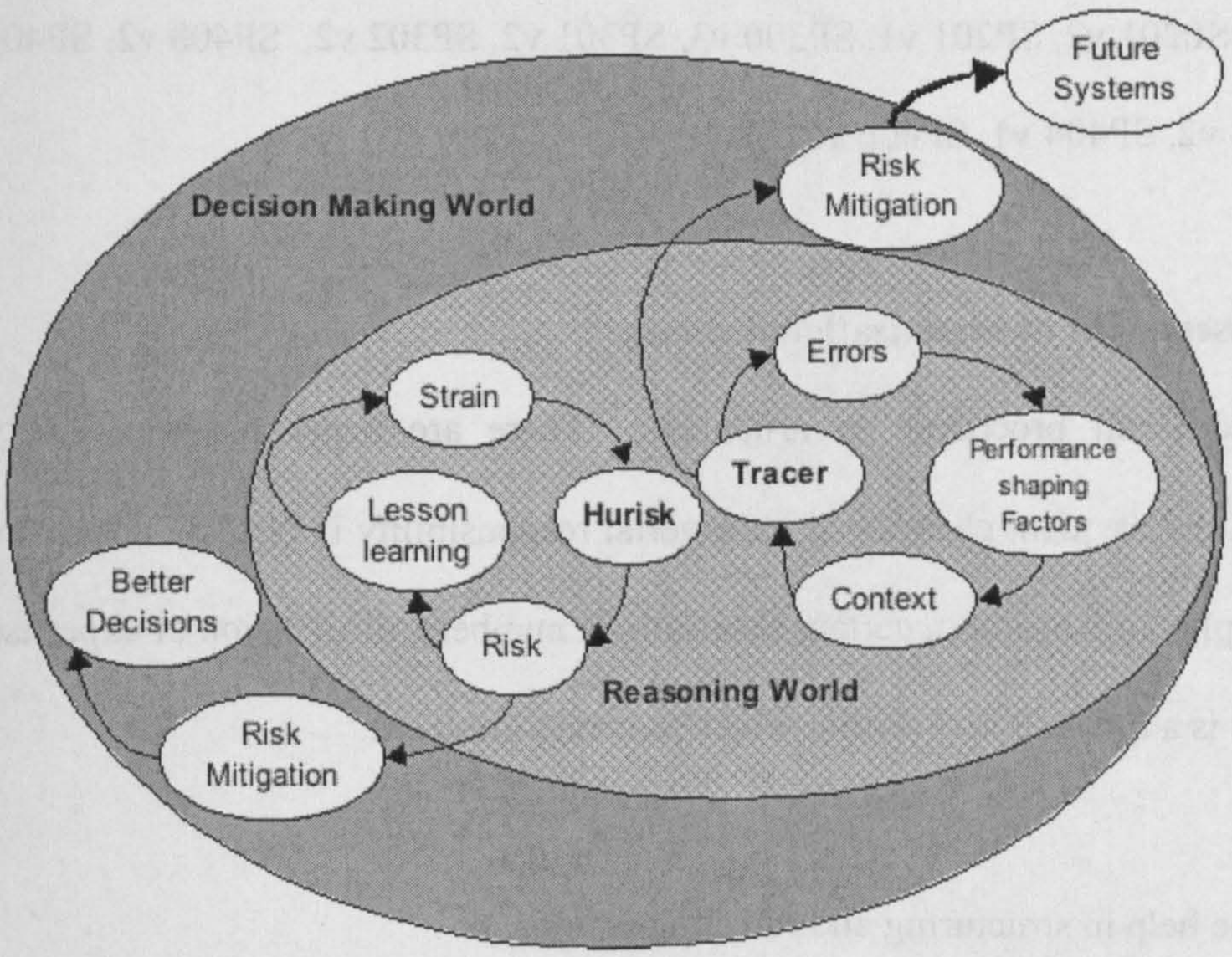
Comparing and combing Tracer and Hurisk

Table 12: Comparison of Tracer and Hurisk

	Tracer	**Hurisk**
Hierarchical Structure	Yes	Yes
Based on a spreadsheet	Yes	Yes
Recording of Data	No	Yes
Allows Subjective Data	Yes	Yes
Contains Objective Data	Possible	Yes
Produces Statistical Scores	No	Yes
Has an Empirical Basis	No	No
Is Retro-fitted to situations	Yes	Yes

The two tools as can be seen do, to an extent, live on a continuum where Hurisk is dealing largely with the organisational phenomena which surround and sometimes make up performance shaping factors and Tracer is looking at the discrete level of the factors themselves and the errors they accompany. In this sense the tools are close relatives and could potentially be of use to the same user communities. Although Tracer does not contain any formal scoring it is, by its very structure, conducive to adopting a scoring strategy like that in Hurisk. If Tracer adopted a more sophisticate scoring structure this would greatly enhance its potential to act as a risk management tool.

Figure 28 : An a priori model of the ATC environment of Tracer and Hurisk



Tracer and Hurisk are both ultimately heuristic tools, Tracer because it produces guided reasoning and Hurisk because it allows the development, monitoring and visualisation of its key objects in a user centric way. The conclusion of this benchmarking study was very interesting for two reasons. First, NATS Human Factors, because of operational and organisational constraints, had resorted to building a heuristic system based on content rich data sources and designed to guide expert judgement. Second, they had stopped short of building a scoring platform for it. Hurisk, by comparison, looked ahead of its time.

Benchmark two: The Safety Monitoring System

NATS has a complex safety monitoring system in place, this is not a surprise. A key benchmarking exercise was to read through the certificated safety statements for the elements of that system and reflect on how Hurisk might contribute something to the overall safety case of a present and future monitoring system. The conclusion was that a fully integrated version

of Hurisk would provide an expansion and enhancement to several safety system components, these were:

SP100 v2, SP200 v1, SP201 v2, SP201 v1, SP300 v3, SP301 v2, SP302 v2, SP400 v2, SP401 v1, SP402 v2, SP 403 v2, SP404 v1, SP406 v2, SP407 v1.

SP100: The safety assessment of organisational change

Currently this is a difficult procedure to articulate. There are many potential safety management issues in the frequent changes of managerial responsibility in NATS. There are also, as with any complex organisation, certain shortfalls in numbers, distribution or expertise at various times. This is a difficult area within which to create guidance.

Hurisk may be of some help in structuring and directing actions.

- It is aimed at assessing changes and potential changes to operations and is able to concretely articulate risks associated with this kind of change
- Hurisk includes communication and management issues
- The lesson learning format may help manage change and "information survival" issues more easily

SP200 Safety Surveys

Most safety surveying necessarily concentrates on the hard evidence of traffic, controlling and technology issues. Surveys tend to deal with these in engineering, physical and statistical ways. The assessment of issues, such as performance shaping factors, which are much more personal and organisational than the human machine and human procedural interfaces, is of growing importance.

Hurisk would be a good tool to use in this case because of its clear focus on such issues and because of its design for continuous assessment and review. It may be able to help assessors make a more well rounded assessment. It could be integrated with other safety surveys throughout the year or act as a stand alone.

SP201 Safety Management System (SMS) Continuous Assessment

The use of tools such as incident reporting, event monitoring, and trends analysis are all essential to keep the SMS in a healthy state. The importance of lesson dissemination has been more widely recognised. Also, the advent of tools such as the technique for retrospective analysis of cognitive error (Tracer) has begun to investigate incidents and human error from a wider human factors perspective.

Hurisk is a tool built on very similar principles to Tracer but which is looking at the wider organisation which shrouds the safety significant events. As such, Hurisk is not only complementary to Tracer but may be a similarly useful addition to the continuous assessment of the efficacy of the SMS. Hurisk also contains lesson learning functionality.

SP202 Safety Reviews

Safety reviews are an important large scale exercise for continuous improvement both of safety data and safety performance. These include the collation of incident investigation outcome data.

Hurisk's main focus is on the higher level features of the safety culture as such it could contribute to safety reviews and broaden the evidence base that they can call on. Hurisk is also a potential candidate for the need recognised to "fill in the gaps" which the current data leaves. Strain data in particular may provide an explanation for some of the changes in the

objective data. If such links can be made then the dissemination of lessons learned from safety reviews will be more thorough.

SP402 Development and assessment of safety related systems

An important goal in ATC is to understand the dynamic and complex relationships between procedures, people who execute them and the equipment they use to do this. In a highly engineered environment like ATC with a large number of technical specialists. This gives a slight bias towards the technical and procedural aspects of this. The people aspect is itself dominated by the more technical elements such as the Human Machine Interface issues. These assessments are intended as the basis for requirement setting.

Hurisk provides a means of linking these elements but with more emphasis on the people and organisational issues. Thus it may provide a good bridge to the more technical methods. Hurisk also has an "options appraisal" functionality. This could be used to assess differences between requirement sets. In these ways Hurisk may support and strengthen the assessment of safety related systems.

SP403 Systems Safety Case

The systems safety case presents evidence, arguments and assumptions to show that system hazards have been identified and controlled; both in engineering and operational areas; and that qualitative and quantitative safety requirements have been met.

Hurisk data is a blend of mainly qualitative judgement data with hard traffic and controller data. In an operational setting where traditionally quantitative data has been the most functional approach, Hurisk represents a validated opportunity to strengthen and customise the qualitative basis of safety case data.

SP404 Systemic Safety

To present safety evidence and arguments through a systematic approach to safety management, to justify the claim that the operational safety of a unit or en route facility is adequate for its role.

Hurisk is a long term recording system with an audit strand which allows reviewing over selected time periods. As it is a reasoning platform it shows how evidence has been balanced over time.

SP407 Safety assurance

To ensure that NATS has adequate safety assurance for the equipment elements of the ATC systems that are procured under contract and used in the provision of air traffic service.

Hurisk has a set of tools which are designed to look specifically at projects. These have been based on a typical procurement and design paradigm to allow the identification of strain and risk in the operational setting directly caused by these processes and the resulting technology. Hurisk is very well placed to provide a complimentary data set to existing technical quality control and risk assurance methods. This is because it is oriented towards capturing the human judgement data which is lost in technical representations but contains a very important strand of knowledge which can be transferred to prevent technology transfer failure and costly late re-design.

3.8 Final consideration of results

Risk Management

The National Air Traffic Services, as evidenced in its sponsorship of these projects, understood very well that the safety of the UK airspace was a complex and dynamic thing. This led to the growing understanding, in a burgeoning air travel society, that it was not just the decisions made by air traffic controllers which effected the safety of passengers. The decisions made by engineering designers and operational air traffic control management could conceivably affect and effect that safety. The worry was that these decisions were more invisible, given that they were not scrupulously reviewed and reported. Thus, slowly these decisions might conspire to undermine engineered safety. This opened up the challenge that decision maker attributes should also be modelled because these were, in an of themselves, potential arbiters of a new form of risk.

Having understood that a "human risk" to air traffic was a real potential, NATS wanted to measure it. It is a part of their culture that all risks are to be minimised to generate safety for passengers and protection for property. The "socio-technical system" of NATS therefore was to be the proving ground for a risk model which could compliment NATS' already well developed technical risk controls.

Not 'Human Factors' per se

The human risks in question were not the kind that come naturally from complex technology and human operators. NATS employed a team of Human Factors specialists to assess technology risks. Those risk assessment processes were governed by the strictest technical and legal stipulations encased in The Safety Case. This was something which all NATS systems had to have. Technical risks more traditionally associated with this industry, e.g. the reliability of air traffic control equipment or the prediction of human cognitive error patterns,

were already covered by formal human factors methods e.g. Hierarchical Task Analysis (HTA), Failure Modes Effects and Criticalities Analysis (FMECA) and so on. These were not to be the remit of the Hurisk study.

Decision Making

The decision making covered by Hurisk was not the kind which is carried out by air traffic controllers in achieving the "safe and expeditious" movement of aircraft. Rather, it related to reasoning about risks of all kinds in engineering project and operational management meetings. That reasoning of course underpinned formal air traffic control that in terms of resource, infrastructure, organisation and technological spend. The potential value of a risk reasoning system specially designed for management decisions would be better and more consistent (and transparent) decision making about risk by project managers and operations managers.

The decision making of these managers, it was recognised from the beginning, was taking place in a context of increasing pressure for economy and efficiency of operation coupled to a heretofore unparalleled expansion of air travel in the UK. Engineering project decisions and the decisions which dictate the resources and support for air traffic controllers, as well as changes to their roles, technology and working environments, were being realised in a faster, more dynamic and more uncertain environment. The air traffic controllers' role, as a slow paced and stable structural underpinning of the safety of air traffic management in the U.K., was recognised as under threat. NATS wished to understand that threat and restore balance. The forum for these sort of risks and strains, and their associated decision making, were predominantly discussion and consensus forum, that is management meetings. That fact raised several cultural, organisational and psychological challenges. Hurisk was commissioned with an open-ended remit to explore the concept of "human centred risk" (as

opposed to technology, or human technology interface centred risk) in these forum with a view to gaining more sophisticated and more explicit levels of risk control.

3.8.1 Key attributes of this system

This discussion will go on to consider in more depth the Hurisk 2 development project. Having described key technical outputs in the main body of the chapter space will now be give to discussing four areas:

1. What have I concluded that 'strains' are and how do they work in a system like this?
2. What have I defined 'risks' to be and how do they work in a system like this?
3. Summary conclusions on the concept of lesson learning tools
4. Conclusions on the whole project

What are strains?

It is important to understand for all the positive use of maths displayed here that strain is a completely heuristic idea. We don't know it really exists in the form we are attributing to it and therefore it remains unclear how to measure it. What we are measuring therefore is a proxy. It is still a proxy which needs rules. Hurisk works on the basis that any single individual strain has the same potential to upset safety as any other. This assumption underpins the heuristic the system.

Strains, simply put, are the operational effects left behind by a bad design process or poor decision making and compromises. Strains are also the operational effects of the instability caused by projects being undertaken, engineering or operational organisation. Strains are not a safety loss, the certificated safety of the system is not, according to standard metrics, reduced. It follows that, if strains exist, the standard metrics are not sensitive to them. This is a key concept.

How do strains work?

Strains are tiny 'ghosts in the machine' and they threaten safety in one of two ways. By being present in large numbers (a critical mass argument) or by working unnoticed over a period time (an attrition argument). The key to strains is that they are a force against safety, they make the air traffic controller's job a tiny bit harder or less logical, they make the organisation of air traffic a tiny bit more complex, they are present in technology solutions which are a tiny bit sub-optimal, they create decision pressures which make people agree with a decision even when they are uncomfortable with some small level of compromise. Also, as the history of industrial disasters shows, strains, or more generically risks, are sometimes tiny but can be expressed forcefully when the engineered system itself is in a unique or anomalous state.

Differentiating engineering and operational strains

Although strains are heuristic, measurements were developed for three kinds in the NATS organisation because these could be seen to be working differently. Thus there is a specific form of strain caused directly by instability around large engineering projects. This was called "engineering strain". There is also a form of strain on the organisation's decision makers. This is coming from the social and economic forces of operating competitive, but safe, air traffic control services. This form of strain can be generic, because the operational health can vary. This form can also be specific to a large project or programme (not all significant projects in NATS are engineering focussed). These were called operational strains and therefore came in generic and specific forms.

Strain analysis for air traffic projects and management discussions now has a detailed methodology to it, a fixed rationale (based in applied research) and a suite of on line reasoning visualisations to help track strain at varying degrees of detail within and between

users. This approach because its content is embedded in real descriptive concerns, compliments the more focussed technical safety case.

Strains give a heuristic system which nonetheless preserves necessary assumptions e.g. equality of strains. It comes with a highly structured measurement system which is designed in two formats offering basic high-level impression or a comprehensive analysis. Importantly these fit together interchangeably, as appropriate, in providing the final strain summary.

Strain measurement now has a highly structured questionnaire basis. The questionnaire creates a 'journey system' for rationality. The scores are indeed relevant, but the journey to the output is actually about the reasoning stimulated by thinking in a calibrated way about the questions (a key feedback from user trials). The scores create a proxy for that rationality and introduce a second level of economy in future scoring (adjusting from a benchmark) and reasoning (being able to review why was this score given, is it still relevant).

Strain consideration forces users away from very open-ended processes such as discussion. The tool does this by causing a pre-analysis activity and offering the capability to see a raft of focussed comparisons between and within groups and individuals over time. The comparisons all have high cultural value within the organisation (again backed up by user trials) because they came from the culture, and are written in its language. The application of this tool, particularly in group reasoning tasks seeks to expose the underlying rationality that drives preferences. These drive decisions which of course in this context drives some of the safety of aircraft and passengers in the sky.

What I have achieved is proof of concept for a new reasoning device, strain. That device is aimed at 'the human element' source of risk in air traffic proven in the first study. I have used

measurements within a parsimonious system to elicit decision maker comparisons based on an agreed underlying model of the world which can be used in real time comparisons to aid rationality. This model is in the reference grammar of the decision makers themselves even though it is using mathematical modelling to produce decision support inputs.

What are risks in Hurisk?

Risks are simplified for the purposes of Hurisk. Simply put, they are a summarisation of negative events which may occur to upset safe operational functioning or the safe execution of a project. NATS already uses this concept of discrete risks in its safety cases.

The idea of working with risk is straightforward.

1. In a safety critical setting it is important to demonstrate that you are cognisant of "the risks" of what you are doing.
2. As a form of communication, phrasing unsatisfactory outcomes as individual risks is a good strategy.
3. If you can define something as a risk then Hurisk allows you to measure it and chart its progress over time and in reaction to control features.

At the descriptive level, a risk in Hurisk is determined by the user. The user provides discrete events which pertain to the assessment question. This is at a level where they can describe them efficiently. These events are things which they feel can or will threaten safe operation or safe project execution. The outworking of these events are likely to be very complex, but Hurisk is being used to keep high-level tabs on them and the way they behave over time. A summary of the guidelines used for writing risks is seven-fold:

1. A risk does not equate to a hazard
2. A risk does not equate to an emergency
3. A risk happens in time

4. A risk is phenomenological (it can actually happen, it is not just a "fear")

5. A risk stipulates a measurement

6. A risk should not be too large

7. A risk should not be too small

Risk 'scenario' generation

It is conceivable that risk tolerance would vary throughout the life of a project. The risk tolerance input described earlier is therefore kept separate from the risk scoring. This allows users to experiment at different times with differing levels of risk tolerance whilst holding all other values constant. Thus the output of Hurisk, the key risks, can be explored for a given risk tolerance scenario.

With this risk tool I have made some important steps forward for an organisation which already has a detailed understanding of technical risk sources and their management. We have created (and had accepted) a bounded definition for "the human element" risk as a new form of risk. This has been accepted as complimentary to not only technical risk but to formal human factors risk. A lot of that success would seem to lie in no small way in the rigour with which this sort of risk can be defended as coherent (ultimately challenging the existing forms) and which has a support structure (in the form of the seven rules) which maintains the force of that definition into use.

When it comes to measurement things are tightly defined and we have produced a complex, but coherent, multi-attribute rating system for this risk. This makes it formidably transparent what the reasoning behind any high or low risk is (again existing systems do not actually have this transparency). In the particular case of project and operational decisions we have moved the organisation a long, long way from consensus based discussions resulting in high, medium

or low risk. Of course, in keeping with the desire to reflect existing heuristics, the system permits this level of coding.

As in the case of strain the measurement the tool is not only powerful, but it is flexible too. This recognises the highly time pressured nature of these environments and is a key to transferring the technology. The target of an hours worth of assessment for a reasonable portfolio of risks has been met. In fact, using the very bluntest form of the risk tool, this has been reduced to ten minutes. The quick and simplified scoring alternatives are seamless in the system with the detailed ones although clearly more fundamentally heuristic (in the sense in which we are using that term).

Also as in the case of strain the risk tool allows for (encourages) a time dynamic assessment in a world which was typified by "snap shot" approaches. Like the strain case, the risk tool has a range of visualisations which are conceptually quite distinct and allow a detailed form of auditing within and between users as well as "impressionistic" comparisons. Risks can be reviewed and re-scored at any time by any user. Analyses results show comparison within and between individuals and groups over time. More powerfully than that however, it is also possible to look at hypothetical shifts in risk tolerance over time and this may prove an aid to refining the action plans around risks.

In view of the range of input modes from simple to complex, and in view of the idea that deeper reasoning is considered of higher quality, the tool contains a quality filter for depth and shelf life of an individual or team's assessment. This is placed in the primary output. Its use is not stipulated as it is an essentially passive recorder, but it is mindful of the very real human dynamics of risk judgement (i.e. that people do not have a lot of time to review scoring) observed in these settings.

The risk tool, importantly for these summary conclusions, produces risk data which is in the same currency as strain data. That synergy is important when one comes to consider, as we will later, that they are really two sides of the same coin.

'Lessons learning'

In the case of developing lesson a learning technology and approach we were, perhaps unsurprisingly, at our most challenged. Lessons are so much more intangible than our highly structured risk and strain concepts. They are inherently more difficult to define and one is faced with the problem of mapping any definition to the appropriate arena. Lessons could be about speculation on facts or indeed on persons. Lessons have no natural currency that is at all approaching anything one might summarise with a mathematical heuristic. Lessons are inherently subjective. Most importantly, lessons are the very essence of professional judgement which is, of course, the ultimate intangible in the safety bearing qualities of a human - technical (and socio-technical) interface.

What our early prototypes reported here do stimulate is some fuller thought as to how this area could go forward. We re-visited and improved the triads from Hurisk one, a simple, but quite elegant concept. What we found was that one could attach very robust scoring methods to this and it created differentiation which could be easily highlighted between decision makers. The utility of having that is not certain, but it is feasible to do and it was an improvement on the notation output that these had previously given.

We were able to create a raft of possible decision-maker and decision-content comparisons. These could work as a sort of decision priming phase. That might be particularly important to keep an eye on upstream safety threats entering the system in a similar manner to that hypothesised in the strains and risk work. Lessons could lead to an understanding of how the decision space and its safety bearing qualities are being conceptualised by different groups. That would clearly be a window on the way people chose to reason and the weight being given to particular classes of decision. This is all very soft at the level we took it to, but clearly not without merit.

We were able to experiment with slightly more sophisticated narrative modelling forms in the shape of the questionnaire and the two forms of a detailed taxonomy of influences. There is clear room demonstrated here (although an agreed model would be a pre-requisite) to probe a good deal more deeply into decision rationale. This sort of analysis can only help shape decision making processes, particularly if these are highly time constrained, which we knew was the case. An off-line analysis that caused decision makers to reason out their positioning beforehand might improve decision making.

Lastly, we looked at projecting this sort of decision taxonomy approach at a model of the "stereotypes" (some might argue archetypes) of decision influences (and influencers) in the NATS environment. This, almost a self-knowledge proposition, threw up some interesting findings. It was possible to highlight deep contradictions in the safety emphasis of a decision space just by placing it under a different stereotype. Although somewhat hypothetical, at the level of evidence we have, this is highly interesting again in understanding the forces which shape safety decision definition.

Benchmarking

Comparing Hurisk with Tracer produced an interesting caveat to the Hurisk 2 project. It could be seen that NATS' Human Factors department were experimenting with heuristic reasoning tools for sensitive or unclear areas of risk. Where Hurisk has the upper hand is in the fact that it is centred on detailed scoring. Tracer's developers were not able to take that step.

Assessing Hurisk against some of NATS' safety standards proved to be a very interesting technology transfer exercise. It is clear that Hurisk can contribute to the levels of safety

which these standards require in new ways which in some cases would actually improve on them.

3.8.2 Where has this research taken us?

The heart of this work has shared a journey of discovery with a large, complex, safety-critical organisation in evolution. At the beginning of that journey NATS "knew risk" and "did risk" as a normal part of their charter. This was both technical and Human Factors (for air traffic controllers) centred. Through sheer diligence to prevent loss of life in air travel NATS encouraged self-doubt and always entertained the possibility that there was more to learn about risk. At the time of my research a number of other parallel projects were running. My area was "human centred risk" particularly that part where upstream decision making (increasingly driven by commercial forces) might somehow impact downstream safety at the operational level in a way that existing alert functions were not sensitive to.

As a fundamentally socio-technical system based on risk NATS was hungry for "a systems solution" as the countermeasure for any new risk source identified. Once we had convincingly identified that there was such a risk source I managed to convince them of the idea that a human centred risk system might best function heuristically. In creative terms this opened up a huge range of options. This debate was therefore fuelled by one grand overarching heuristic: the thermometer and its partner the scale. In sympathy with aviation heritage these were stylised as clock-faced gauges.

NATS wanted to "meter" the risk temperature at the air traffic controller level, at the engineering project level and at the managerial decision level. Importantly, which is why lesson learning was added, they wanted these three communities to communicate with each other and a common tool used by all for different applications seemed an excellent method.

As a result of this research NATS had a web-based safety tool called wHurisk. It has the capability of providing robust decision support in the areas of upstream human risk entering the safety supply chain through management and project decisions. Examples of these human risks have been rigorously researched in the widest possible range of case studies. These case studies cut across the breadth of the safety bearing properties of NATS' operation covering technology, training, organisational behaviour and, above all, decision making. A comprehensive measurement system for such risks has been designed.

wHurisk has the capability of analysing normal operations, operations under project pressure and large engineering projects for attrition of safety they cause. This attrition is in the from the hypothetical entities called strains. These have been thoroughly researched and described and turned into a measurement tool. This tool has gained high levels of acceptance in NATS, particularly with the engineering community. A comprehensive measurement system for such strains has been designed.

wHurisk, in the assessment of the priority strains and risks in any NATS proposition, has the capability of folding inwards and forwards basic lessons to be learned in future similar projects or decisions. The surrounding research has also unearthed some very interesting, if prototypical, ideas for comparing decision spaces around safety.

The working concepts in wHurisk have been validated by a NATS expert steering committee and wider expert groups contributing to concept development workshops. They have also been validated by a sample of "ordinary" NATS personnel utilising the tools in their daily work (not reported here). The thing that is the most interesting about wHurisk is that it is a

heuristic reasoning platform for air traffic safety decisions. Technically it does this in three essential ways.

First, it plays upon the essential natural heuristics of an technically defined organisation like NATS i.e. the use of graphs, charts, gauges, and meters to provide rate and state information about physical systems and "potential conflicts".

Second, it has detailed measures of its core objects on a 1-100 scale and these are combined meaningfully, by some basic mathematics, to give them properties such as summary values.

Third, the way in which measurement takes place is flexible there are interchangeable methods to measure risks and strains. Some are very quick and very dirty, others are very detailed.

Philosophically this is all heuristic reasoning in one overarching way. That is that the things these systems measure (in some detail) are not only impossible to measure, but they are impossible to observe, in the scientific sense. Risk, strains, lessons are collective spectres without physical form or phenomenological rules. What they do is help people to have a shared narrative of their fears associated, in this case, with the physical form and phenomenology of an air crash. The collective anti-thesis to these fears is the controlling grand heuristic of NATS other shared narrative: safety. In modelling and supporting decision making wHurisk, in all likelihood, was making a contribution to NATS' confidence in its own reasoning on safety.

The degree to which these heuristic devices might be improved upon with reference to the, inherently similar but far more formal, systems of subjective Bayes methods will be discussed at the end of this thesis.

The bridge between NATS and Unilever

The similarity between NATS and Unilever is not immediately apparent, it is true. It lay in both being risk users an in one other important factor, Unilever's appetite to better measure risk had led them searching for ideas to use. This was to prove an excellent test bed to export key successes from Hurisk and assess them in a new, and very different, environment.

4. The Unilever Projects

Introduction to chapter 4

By its own estimates Unilever products are used one hundred and fifty million times per day across the world. It is one of the world's largest Fast Moving Consumer Goods Companies. It is operating in eighty five countries and has over two hundred thousand staff. The execution of it's business directly impacts on three million lives.

All large companies face "issues and crises". Most have an issues management system in place to cope. Issues management is increasingly becoming a professional discipline supported by a worldwide network of academia, associations and consultants. With the scales upon which Unilever operates, the timely and effective management of issues is a business critical activity. This is not only to preserve continuity of business. In a world of globalised media, this is also critical to preserve a contradiction free corporate reputation.

Issues management as a context

In early 2005, struggling under a portfolio of nearly one hundred issues, Unilever recognised that it needed a more formal system of issues management to replace its historically devolved systems. The new system was to be, in their language, "a one Unilever system". "One Unilever" was a reference to one of the largest re-organisations the Anglo-Dutch company had ever seen. An unparalleled level of rationalisation and re-organisations of Unilever's global business took place. This included a project to deliver the one Unilever issues management system and all of its associated processes, databases and tools.

Key to the success of this global project to improve issues management was the need to develop a coherent system of issues prioritisation. It was decided that this system would draw directly upon some of the main concepts developed in the Hurisk project. They would be

modified (initially simplified) and modelled to suit this new applied sphere. The new tool was named "Descartes".

The overall aim of Descartes was to provide effective ways for a cross section of decision makers from this industry to come to a shared understanding of Unilever's priorities in order to mobilise resources to manage these. The issues prioritisation system integrated into the issues management process (itself in a fast evolving state) would thus provide a new form of decision support for Unilever's risk management and mitigation.

This research was highly detailed and under high levels of applied organisational constraint. In the interests of the reader the development Descartes is portrayed as three distinct case studies. These are:

1. The development of score-cards for issues prioritisation
2. The development of visualisation environments for issues prioritisation
3. A summary review of heuristic concepts in issues prioritisation

4.1 Summary of Chapter 4

Score cards

Issues prioritisation is defined. The methods of scoring and scoring processes prior to this work are described. The main approach to that scoring (the post it pad exercise) is given a detailed critique and advantages and drawbacks are considered. Advantages include familiarity, simplicity and ease of use. Drawbacks surround sources of bias, inconsistency and the lack of any audit or justification for priority. The development of four separate score-cards for improved rigour in prioritisation, with different user communities, is reviewed and a final set of concepts for practicable improvements to issues scoring proposed.

1. The practice in the HPCE group of using an off-line scoring method based on a spreadsheet is considered and some early improvements are reviewed. The effect of these improvements in terms of the behavioural frame of reference, scale design and simple statistical analysis are briefly examined as a function of heuristic reasoning they support. Continuing limitations in the light of these improvements are also considered.

2. The practice of the ICFE group of using a more sophisticated combination of plenary post it pad and review exercise is examined. Results of a workshop approach to improve the validity of prioritisation rationale and tools are considered. Definition strengthening is demonstrated to be a key route to better reasoning with this group also. They were able to accept more radical improvements to their scoring approach by introducing a multi-attribute scale with a detailed behavioural basis. A more detailed scoring process is also considered. This case showed that significant changes to prioritisation process and content were acceptable and considered effective and that these could be grounded within organisational constraints such as time.

3. In the UIG case new impact measures focussing on "business drivers" were sought using a number of complex scale ideas but the cardinal effect of the need for simplification discounted all but a few of these. A final score-card for the global team created further improvements in the area of scale definition and the use of bi-polar scales allowed the language of "advantages" to be held in tension with that of "threat'.

4. Additionally, this UIG team developed a strategic score-card to address their specifics needs to communicate with senior audiences in the lingua franca of corporate reputation. This totally new strategic score-card was based on an analysis of extant strategy documents.

Visualisations

Users were already plotting the results of their early prioritisations. These plots were reviewed and re-created in a graphical computer programme which linked them directly to the score-cards, also stored in the programme. A first result was a visually improved version of the plots already in use.

Problems of data volume in the ICFE case led to an innovation to allow grouping of issues so that these could be rolled up into single data points. A new visualisation was developed to show the 'hidden' scores of individual items in groups. New functionality was added to allow direct interaction with groups in a dedicated visualisation which supported better reasoning about groups.

The more demanding case of the two UIG score-cards (including advantages and impacts, assessing strategic impact) created visualisation innovations away from the traditional quadrant plot. Time was now also visualised as a linear relationship between now and future data points. The strategic plot was created as a boundary plot to reflect negative and positive

values in both scales simultaneously. When combined with time this was seen as very powerful

Heuristic reasoning

Heuristics found in issues prioritisation surround priority itself, the measurements used and the graphs which represent the data. One case study shows how these could all be modestly improved and the improvements were valued which could be described as working within the essential natural heuristic on offer. One case study shows how more radical changes to the heuristics were also possible and accepted, this could be described as expanding the essential natural heuristics. A third case study shows that innovation in the use of heuristics was also possible such as complex grouping techniques or condensing large supply chains of scoring information. A final issues prioritisation process benefiting from all of these developments was successfully rolled out in Unilever.

4.1.1 Some key points from this chapter

Score-card development

- The main technique observed for issues prioritisation was the group post it pad exercise, this is a highly limited approach and prone to inconsistent reasoning.
- The simple and expedient scales of priority currently in use were not considered to carry enough rationality into later decision making but a system with usability and flexibility was highly desirable.
- It was possible to improve the essential natural heuristics of existing reasoning by small improvements focussed on better attribute definition and improved scoring and simple statistics

- It was possible to expand the essential natural heuristics of existing reasoning by more sophisticated improvements including the introduction of multi-attribute scaling and flexible forms of data reduction through grouping.
- It was possible to develop measurement heuristics for issues which focussed on advantages as well as threats.
- It was possible, at the behest of strategic decision makers, to adapt Unilever's company strategy into a score-card.

Visualisation development

- Visualisations improving upon those already in use were easily accepted, with these came the ability to show more data in a more sophisticated way.
- Problems of data volume led to functionality to group issues into data reducing clusters which greatly improved the reasoning power of the approach
- Despite an oft stated desire for simplicity the users were able to accept and use complex visualisations involving larger and more detailed plots
- The introduction of time modelling was a valued breakthrough in the utility of this kind of visualisation to assist reasoning
- Entirely new forms visualisation which broke previous conventions were readily accepted.

The development of heuristic concepts

- It has been possible to observe and develop the heuristics which are in use in these sorts of settings.
- These forms of reasoning can be improved, revised and innovated. In all cases a high degree of technology transfer was observable.

4.2 Case study one: Score-cards for issues prioritisation.

Introduction

Issues prioritisation is the discipline of examining multiple threats and opportunities to ones business over time. An issues prioritisation scoring method can either be informal e.g. group discussion and voting; semi-formal, such as post-it pad exercises; or completely formal such as utilising a validated score-card. Issues prioritisation scoring is one aspect of a wider issues prioritisation and management process.

This case study tracks the development of four formal issues score-cards in Unilever under the supervision of the Unilever Issues Group (hereafter UIG) was a group. This oversaw the global governance of issues management. The development took place with the assistance of Unilever's issue managers and their expert teams and that of the Global Issues Manager the single expert at the head of the issues management community and who trains, supports and organises them. The aim of the development was to support more efficacious methods to manage Unilever's issues portfolio.

Whilst the global issues management process was overseen in the way described above, a number of similar, but smaller, structures were replicated in different Unilever Business Units. Two of these units were Ice Cream / Frozen Foods (ICFE) and Home and Personal Care (HPCE). These units developed score-cards for issues prioritisation ahead of the UIG's efforts. These score-cards were early templates.

4.2.1 Aims

This study describes the applied development of four issues prioritisation score-cards, that work can be summarised under the following aims.

1. To work with subject matter experts to extract user requirements

2. To establish a critique of the current best practice, from a process and technological point of view, and ascertain the aspiration for future process and technology.
3. To design and evaluate a series of prototype score-cards.

4.2.2 Objectives

Three projects to deliver four score-cards to different user groups ran in a sequence. The projects can be described by the following (sometimes overlapping) objectives:

1. To assess existing practices in issue prioritisation scoring
2. To develop the user requirements for new ideas for issues prioritisation.
3. To derive separate, business unit specific, score-card contents.
4. To evaluate potential score-card elements with users and commissioners.
5. Internal testing and validation of score cards.
6. Iteration of score-card design.
7. External testing, validation and final improvements to score-cards.

4.2.3 Methodology

The following methods were used (to differing degrees) in the development exercises.

- Observation of existing practices in issue prioritisation scoring through attendance at project meetings and discussion with project users.
- Facilitated group workshops with end users to develop content concepts and scoring.
- Structured group discussions with commissioners to assess the suitability of business unit specific score-card content.
- Structured and ad-hoc one to one interviews with users to evaluate candidate score-card elements and assess these for consistency and comprehensibility.
- Content analysis of Strategy Into Action plans as a basis for score-cards.
- Test of score-cards on samples of issues in plenary discussion.

- One to one interviews with issues managers.
- Final field testing of score-cards including comparison of group scoring and off-line individual methods.

4.3 Consideration of Results

Review of existing practices: The post-it pad exercise

The original post it note was made by the 3M Group and was launched in 1980. These yellow sticky note pads would later give their name to a group reasoning exercise: the post-it exercise. Variations around that exercise were observed and reported at all levels of Unilever's business up to and including the executive.

To better understand the requirements for issues prioritisation score-cards five post it pad exercises were observed. The flow of the exercise and the main steps were analysed to assess what was being achieved and in particular how data on issue priority was being conceptualised and managed. The following description gives a typical example and summarises pertinent results of observations.

A facilitator would, on an A1 flip chart, either:

1. Draw a set of x and y axis, or
2. Divide the page into four equal quadrants.

The 'x and y axes' of this drawing are labelled. Most commonly, for prioritisation tasks observed, the labels were "Probability" and "Impact".

The subsequent stages of this extremely soft approach, and the order found in the following description, should be thought of as typical, not definitive. The titles of the stages are a

necessary way to describe observable phenomena but they were not used by participants. The titles are not meant to convey that this technique has formally been described in any peer-reviewed literature or that the titles themselves refer to known reviewed techniques.

a. Generation In the group form of the exercise each participant is given a set of post it pads and a pen. They are charged to write the names of the current issues which they are aware of and then stick these to the flip chart thus indicating their relative positions on the two axes.

b. Consideration Once all of the participants have completed this the group gathers around the results to read them and ingest what they are saying.

c. Disputation The facilitator will begin by having an open discussion on agreement and disagreement with the relative positions of all of the post-its and the group will begin to debate relevance and position of particular notes.

d. Clustering As a result of the discussion, or as a formal instruction from the facilitator, the participants will begin to cluster similar objects together by sticking the post-it pads on top of each other. Note this essentially is over-writing the position data of those notes which are moved. When clustering is in full swing it is observable that several small teams, or individuals, may take control of the board simultaneously. In more structured exercises only the facilitator moves any notes and this only by consensus.

When clusters are agreed it is common to replace the clustered notes with a single note. Sometimes this adopts a name which is common to those it replaces, sometimes it also has a bulleted list of the original names under a new title.

e. Re-scoring Once the definitive set of notes has been formally or informally defined, the relative positions of the notes is often tailored by making pair-wise comparisons between notes which highlight any anomalies in their positions.

f. Consolidation As the exercise is brought to a close the picture consolidates, clusters are accepted, further combined or spliced back into smaller units. New items are sometimes brought to the chart as a result of omission or ideas generated from seeing the "mature" chart.

g. Reviewing The rule of thumb tended to be that the upper right hand quadrant of the page was the area of the highest priority/interest and the lower left the lowest. The group would at some point review a "whole picture" on this basis discussing the relative priorities. Once this was pronounced as satisfactory the chart would be frozen.

Some possible benefits of the post it pad approach

- It is quick and easy to do and doesn't require any complex equipment
- It is familiar and expected, people are comfortable to do it for a range of problems
- Its output is easy to use because of the aforementioned familiarity
- The results are uncomplicated fit into a small space

Some problems with the post-pad approach

- Known psychological biases of group discourse such as the effect of hierarchy and assertiveness, lead to some opinions carrying more weight and agreement
- Anchoring effects are potentially introduced very early on where later variables are scored relative to existing variables
- Position of a post it relative to an A1 sheet is a poor proxy for a real measurement scale (and heavily mediated by which size Post-it note is used)

- The rationale (and variability) for the prioritisation decision is not formally captured
- The clustering approach is neither formal nor criteria led
- The analysis is bound to the time at which it is done, essentially being a snap shot of the current state of affairs

Challenges for improvement

For issues prioritisation in the use of the post it pad exercise, or other similar data reduction techniques which provide quick and malleable processes to discuss complex problems, the key weakness is the rationale for priority. The necessary justification of the priority decisions was not observed to be carried forward in any systematic way.

When one prioritises a global issue in a modern industrial setting the money and resource allocation to that issue is significant. Interview data suggested that the acceptability of simple rank order prioritisations, or priority decisions without any justification or supporting data, was being challenged. The task of producing the results of a post it pad exercise was seen to introduce post rationalisations or outright bias. The use of the outputs of these kinds of exercises was being perceived as highly politicised and tantamount to a complex form of lobbying, not measurement.

Discussion of post it pad with users

The sorts of measurements of priority currently in use ('1,2,3' or high, medium, low values) were not considered sufficient to support decision making. Prioritising in this rough and ready manner also created at least three other associated problems. First, checking the rationality of why one object is prioritised over another requires you to re-visit the discussion with the expert. This was so highly inefficient as to never happen. Second, because of the vague criteria for discrimination priority, the Post-it pad exercise was essentially another form

of "lobbying tool", in essence little more than a physical capture of a discussion. Undoubtedly the processes like the post it exercise did facilitate discussion, but whether it really came to any conclusions that a round table discussion could not was called into question. Third, whilst it appeared to be more systematic, the approach of the post it pad exercise was not reducing the known biases of discussion processes.

The retort to such challenges of course was that there was no better system with the same usability and flexibility available for busy people to use.

4.3.1 A first score-card (HPCE category)

The HPCE group had attempted improvements to the measurements in their version of the post it pad exercise. HPCE had a geographically very dispersed team so this exercise was a mix of opportunistic group activity (e.g. at meetings) and off-line scoring via a spreadsheet. Team members in the off-line exercise made a 'high, medium, low' type judgements on two attributes: probability and impact, (these an obvious legacy from early risk plots). A qualitative variable, called 'hot or not", was also measured. This data was collected by e-mail. The team results were then aggregated by the process owner and a quadrants plot was produced as a communication of the results.

Two measurement improvements on the basic Post-it exercise are available here:

1. There was an attempt to create a scale of measurement, albeit a very simple one.
2. There was a second order of data comparison added in a qualitative data point.

Two things are important to note in terms of the use of the approach:

1. The group were scoring independently and using software to do so (Excel).
2. The quadrant plot was only produced once the results from the scoring had been mathematically aggregated. *(This is not strictly true as firm evidence of biasing in interpretation could be observed, e.g. the person responsible for the final output would discount outliers based on "character judgement" , but in principle this was what the users thought was happening).*

Testing improvements

Working with the Issues Manager for HPCE some changes were proposed and tested in the next available exercise, these were:

- Rejecting the "probability" measure and replacing it with a measure of "perception"
- Defining the axes more systematically
- Using a sliding scale (rather than a high, medium, low estimate) with some clear "behavioural anchors" at the poles

- Improving the qualitative variable
- Aggregating the scores by first using simple descriptive statistics
- Labelling the quadrants of the plot

Defining the axes

The text below is an excerpt from the instructions on a revised spreadsheet:

"**What is Business Impact**: Business impact is the degree to which an issue, if uncontrolled, would damage Unilever's business.

You will be asked to rate business impact using a relative scale:

Low = Low impact which is localised

High = High impact, ball-park of 200 Million Euros

What is Perception? Perception is the trickier of the two attributes, it relates primarily to how, and by whom, the Business impact of the issue will be perceived. This is combined with existing perceptions of products, ingredients and processes.

You will be asked to rate perception using a relative scale:

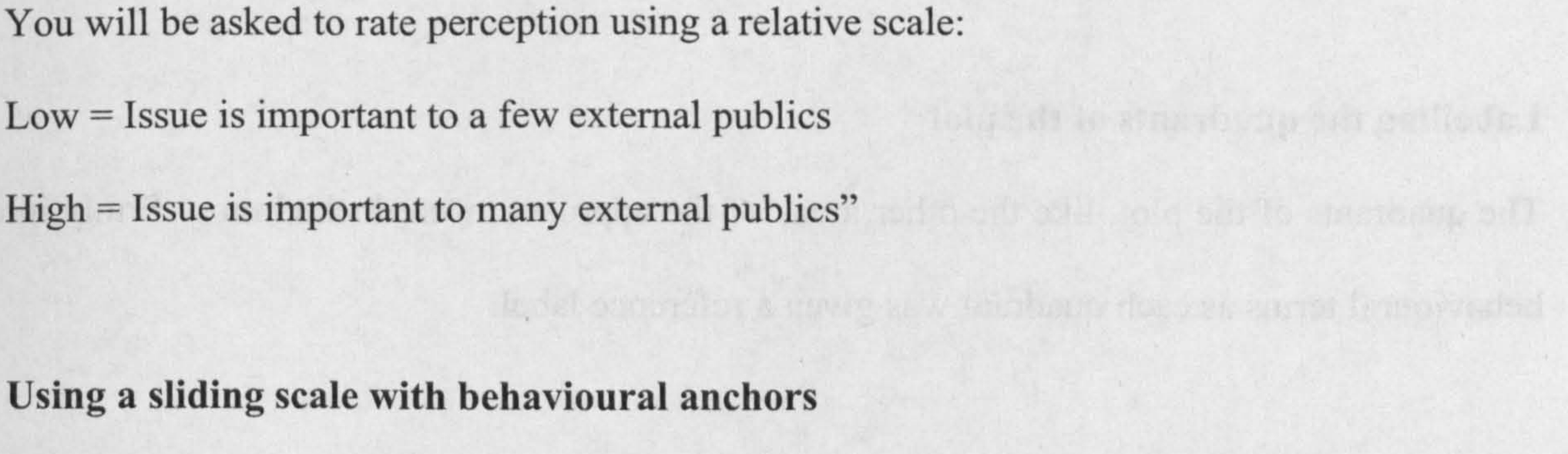
Low = Issue is important to a few external publics

High = Issue is important to many external publics"

Using a sliding scale with behavioural anchors

The scales were presented as follows for each of the issues in the portfolio, note that a nominal figure in Euros was added to the business impact scale.

Figure 29: Spreadsheet scoring tool



Improving qualitative data

A 'hot or not" variable was added, a robust idea to help the category decided whether there were any crises brewing in particular countries. We decided that this too could be sharper by making it a behavioural judgement. Hot or not was retained as the title but the question became a choice between whether there was "current activity" on this issue "live and moving" or "little or no activity". A second qualitative variable was also added to the visualisation entitled "resource allocation". This helped improve understanding of priority in another way.

Aggregating the scores

In the place of using a simple democratic average (i.e. how many of the participants had scored either high, medium or low), mean and standard deviation measures were introduced into the analysis. The mean was used to create the aggregate position. The standard deviation was converted to a 'disagreement index' around the business impact variable and this was translated into a qualitative high, medium or low.

Labelling the quadrants of the plot

The quadrants of the plot, like the other areas of the approach, were locked more firmly into behavioural terms as each quadrant was given a reference label.

Results of these improvements

The HPCE category of Unilever were engaged in an important measurement exercise. From it they wanted to be able to prioritise the issues in their portfolio in an agreed manner. Whilst they had adapted the Post-it pad exercise for this, they still retained its essential flavour. Therefore they retained its essential failings also. Following observation of the effect of these improvements (which were accepted as improvements). Some applied principles can be derived:

Improved definition: The frame of reference for measurement (i.e. the meaning of the axes), the measurement unit itself (i.e. the scales) and the qualitative characterisation (i.e. the interpreted meaning of the data points) have all been improved in this case. These have caused the team to work with behaviourally based judgements over ephemeral ones. Importantly, the behaviours in question are still in the natural language of their business.

Improved Scaling: There will be more discussion on the use of scales in this kind of application. What is important to note here is not so much that users went from a categorical (high, medium, low) assessment of their issues to a continuous scale (in fact 1 – 1000), but that they did so happily. The deliberate choice not to display any numbers on the sliding scales, but simply make them positioned on a line, did not upset these users in the slightest (even though this was predicted). Note here we see a slight break with the Bayes paradigm which would use a nominal scale as opposed to an ordinal.

Improved Descriptive Statistics: Rigour, in these settings, is not a mathematical concept per se, but a psychological one. However, the improvements which came about by confronting the individuals to improve the meaning of their measures and to use numerical scales, led to an instant improvement in the applicability of very simple summary statistics to these problems.

Improved Heuristics: It is important to note that this case study demonstrates is that you can work within the 'essential natural heuristics' of a decision making group and improve them. The scoring approach, the analysis and the communication of the results were all in essence only improved, they were not radically altered from what the users were trying to achieve.

Some remaining problems at this level

These improvements increased the overall rigour and had been accepted by users. However, they were only modest improvements and still left behind deficiencies in the scaling an in the problem reduction. This reduced therefore in the validity of the meaning of the results.

Scaling: A complex business proposition was still being assessed on a two uni-attribute scales. The scales themselves related to business impact and perception of an issue. Such complex concepts should really be reasoned about in more detail.

Problem reduction: It is arguable that even with better scaling and a basic statistical treatment, the problem of prioritisation is still being reduced to a core proposition which examines a quadrant plot and asks: "which of these four categories does the issue belong to?" Judgements on variables of this kind will always be essentially heuristic but improving the actual meaning of the results was still a concern. Issues themselves, by their very nature, are highly dynamic things but this assessment produced a relatively static representation of them.

4.3.2 A second score-card (embedding a multi-attribute measure)

Unilever's Ice-Cream and Frozen Foods category (ICFE), like their colleagues in HPCE, managed a portfolio of issues. Their approach had been to use a group Post-it pad exercise and then refine this data in discussion. Their own report on the results of this exercise was a combined chart showing all the data on a probability impact plot. This had a lot in common with the HPCE approach (using probability – impact, using low – high scaling). The drawbacks already noted were also in evidence therefore.

Workshop to improve priority score-card

A workshop with key personnel from the category was convened to examine the area fully. Understanding what their priority measurements actually meant was a key concern to this group as there was a stronger link between priority measurement outcomes and business interventions in this case. The result was that the objectives of the prioritisation exercise were more fully articulated and the definition of its components was systematically revised.

Defining "an issue" and its prioritisation

ICFE arrived at the following definitions:

Issue: "A threat, specific to Quality Assurance or Safety, Health and the Environment, which has the potential to impact (positively or negatively) on ICFE's business."

Issues prioritisation: "is a process which should enable ICFE to: identify; prioritise; communicate; act upon and assess the status of its issues regularly. The approach must be sustainable and reproducible in the future."

Adopting a multi-attribute approach

Probability and impact, were replaced by 'perception and business impact' as in the HPCE case. Unlike the HPCE case however, a deeper set of discriminating variables were worked

up to support their definition. From the available list of variables discussed at the workshop perception would focus on a shortlist of attributes which were meaningful to the business environment and business impact would adopt a heuristic based on Unilever's supply chain these are defined below and were:

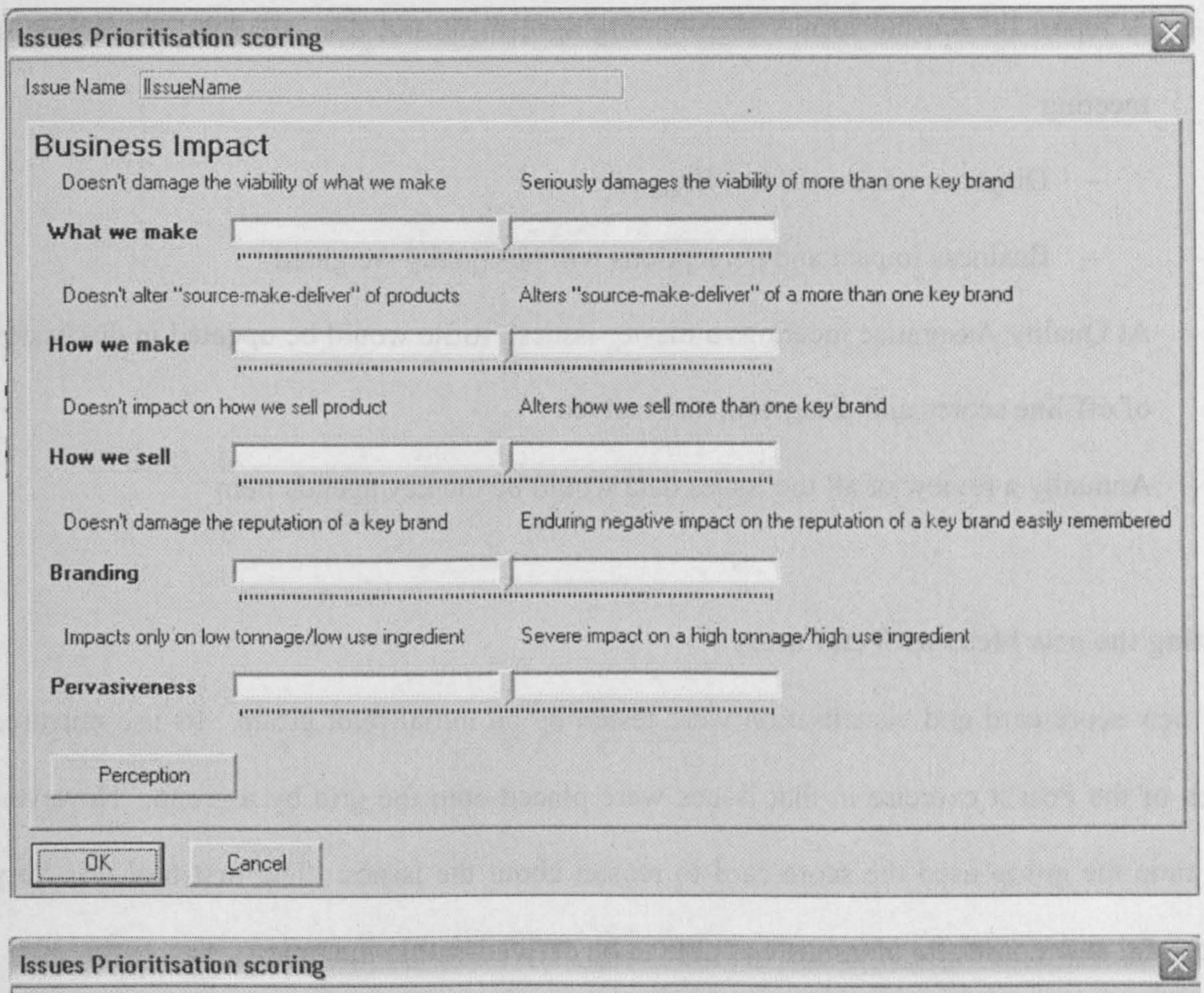
Perception	**Business Impact**
– Intrinsic value	– What we make
– Visibility	– How it's made
– Business damage	– How it's sold
– Ownership	– How much of it
	– What brands are involved

Each of these ideas was translated into a simple question giving nine attributes as follows:

- **Perceived Intrinsic Value**: Is the product or ingredient highly valued by society/ individuals and will this therefore negate the issues with it?
- **Perceived Visibility**: How visible is this issue and to whom?
- **Perceived Business Damage**: How much do people, who are not close to Unilever, automatically think the issue will damage us?
- **Perceived Proportion of Ownership**: How much is Unilever seen as the main owner of this issue?
- **What we make:** Does it damage/ attack the viability of something that we make
- **How we make it:** Does it damage/ attack/ alter the way we make something
- **How we sell:** Does it impact on the way we currently sell products/ where sold
- **Pervasiveness:** Does it impact a high tonnage ingredient/ a high use ingredient
- **Branding:** Does it impact on one of Unilever's key brands

For each of these attributes a bi-polar set of behavioural anchors was worked up. The resultant score-card as it looked in one of the software prototypes is shown:

Figure 33: early score-cards



A process

In the HPCE case a better scoring method had been added to the same scoring process. In this case a new process for prioritisation was also designed. Plenary Post-it Pad exercises were replaced by:

- Off-line scoring prior to Quality Assurance meetings by all team members
- A report on scoring results highlighting agreement and disagreement in advance of meeting
    - Disputed issues will be "flagged"
    - Business impact and perceptions will be equally weighted
- At Quality Assurance meetings a master issues profile would be updated in discussion of off-line scores and disagreement resolved
- Annually a review of all the issues data would be the key agenda item

Testing the new ideas with end users

The new score-card and visualisation were tested by an initial pilot group. Its use mirrored much of the Post it exercise in that issues were placed onto the grid by a group. However, this time the group used the score-card to reason about the issues. The first trial was very successful and a portfolio of issues was able to be derived within the time.

Improvements brought about by this case

This case is a more sophisticated argument than the HPCE case. The rationale, the core technique, the software and the reasoning model all underwent improvement. The result was:

- Improved definition
- Increase rigour of approach
- Increased measurement sophistication

Improved definition: The previous unstructured approach was replaced by an improved set of definitions. The ephemeral notion of 'issue' and the loose process of prioritising have tightened. This has led to increased understanding of the group's intentions within, this process.

Increased rigour of process: The rigour of issues prioritisation process created measurement exercise which was more demanding. "Technology transfer failure" became a concern. A perfectly workable tool might fail to embed in the organisation because it would not be workable within the time available to the process. How long this analysis took had to be "owned" by the users and not imposed by a system. An optimisation approach made this time constraint explicit. The analysis had to take no more than one hour.

Increased measurement sophistication: The existing measurement and reasoning of a professional group (and its heuristics) could and should serve as the platform for better measurement and reasoning. A particular advance made in this early example with ICFE was to create enthusiasm around more sophisticated measurement. In a business environment very much hooked on "traffic lights" and measurements no more sophisticated than "high, medium, low" this example was a major victory. The team addressed the fundamental scaling problems of Unilever's prioritisation approach i.e. the use of only two uni-attribute scales to convey issue complexity meaningfully. The idea of using a multi-attribute score was successfully developed and deployed in a live case.

Conclusions

This group of prioritisation users were successfully taken from a fairly institutionalised Post it pad exercise, one which they struggled to remember the meaning of, to a data driven audit trail of reasoning. All areas of their prioritisation exercise underwent significant improvements in detail and rigour of application. Their reasoning became more systematic and transparent in their own eyes. Importantly it could be applied later to justify the prioritisation of one set of issues over another. This case shows is that a much more sophisticated score-card can be tailored around existing poor quality heuristics and still fit in with the organisational constraints, such as a time limit, which mediated them.

4.3.3 A global issues prioritisation score-card

Developing a score-card for the issues which Unilever as a company were managing on a global scale was a larger challenge. The subject matter was more complex and the scrutiny of and constraints over the end result was greater.

The existing global model for prioritisation largely conformed to the kinds of starting point which had been observed in the two previous cases:

- Application of a "probability & impact" style measurement
- Use of only 'high – medium – low' as a measure
- A quadrant plot as the visualisation
- The use of post it pads and flip charts as the "primary technology"

HPCE and ICFE had been articulating predominantly technical issues in their score-cards. The global company scale required a more detailed user requirements capture process to create a 'scorecard' which could applied to both technical and non-technical issues. There was, as evidence of an ongoing cultural shift in the company, a desire to link priority to "key business drivers" and not just threats. Content analysis of the existing material in this area suggested four fertile areas for further examination:

- Impact on Unilever's corporate reputation
- Actual business impact
- Control
- Perception

a. Impact on Unilever's corporate reputation

Three groups of variables were viewed as crucial: Unilever itself, Non Governmental Organisations and the media. The desired score-card should reflect these in some way.

Table 13: Corporate reputation early ideas

Variable	Measure or comment
Subject of current NGO campaign	Highlight likelihood and credibility with NGO/severity issues
Conflict with corporate social responsibility	Highlight actual or perceived and severity
Conflict with code of business principles	Highlight actual or perceived and detailed severity
Front page coverage in the next 4 months	Contextual statement about likelihood and nature of story

b. Actual Business Impact

A simple measure of business impact avoiding a complex costs benefits assessment was highly desirable. Trade off between the tangible i.e. costs and the intangible i.e. brand reputation loss needed to be understood if possible. Note the very rich, if inconclusive, nature of the possibilities for encapsulating impacts.

Table 14: Competitive advantages ideas

Variable	Measure or comment		
Actual Business Impact	Mainly financial context	Assessing in a quantitative or qualitative way what the impact to Unilever will be if the issue continues unmanaged	Financial or financial equivalence
Cost of issue current impacts	Financial estimate	several multiple choices about estimate of cost	
Cost to put right	Financial estimate	green book approach, economist to help, prototypes/benchmarks approach	
Or use variation on impact of what we make, how we make it, how we sell it, where we make it	Slider bars against severity of impact against a "blood chilling" cost extreme	handles on subjective reasoning benchmark from unilever examples	
Impact on key brands	Mainly descriptive data	Understanding the nature of the behaviour of the issue as it relates to the understanding of brands by producers, consumers and shareholders	Described threat to named brand with associated financial estimate
does it touch the foods 9 or hpce 6	how many of these does it impact and how hard	when do we go alone and when do we go with competitors	
Reputation	Qualitative scale for severity and recoverability		
Efficacy package	Narrative detail of which actual aspects of brand efficacy will be threatened and in what way		

c. Control

Assessment of the controllability of risks again took the form of a sort of costs benefits assessment, this time about likely controls.

Table 15: Influence and control ideas

Variable	Measure or comment		
Control	Behaviourally Anchored Rating Scale and Q&A	Quantifying Unilever's strength in the debate and decision making arena	Actions and costs balance
Ability to Influence	Estimate of how much the issue is being or can be defined by us	"Is it worth it index"	
Debate control	Estimation of how much of the discussion on the issue is controlled by us	What are the risks associated with it	
Complexity to control	An overall picture of the "issue size" in terms of the cost to control if one wished to do so in relation to numbers of discrete players, profile and perception of expertise authority and trust		

d. Perception

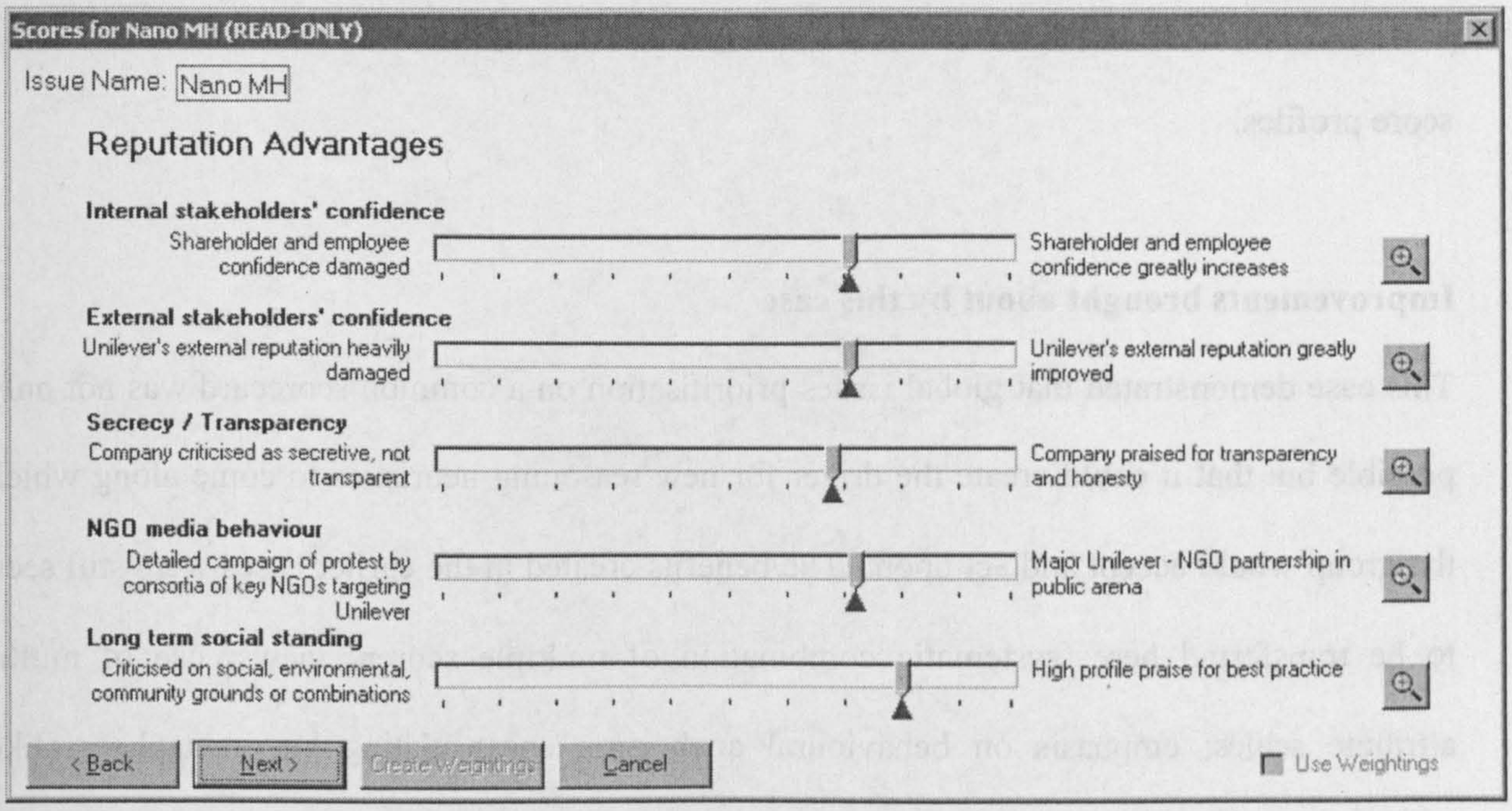
Table 16: Perception ideas

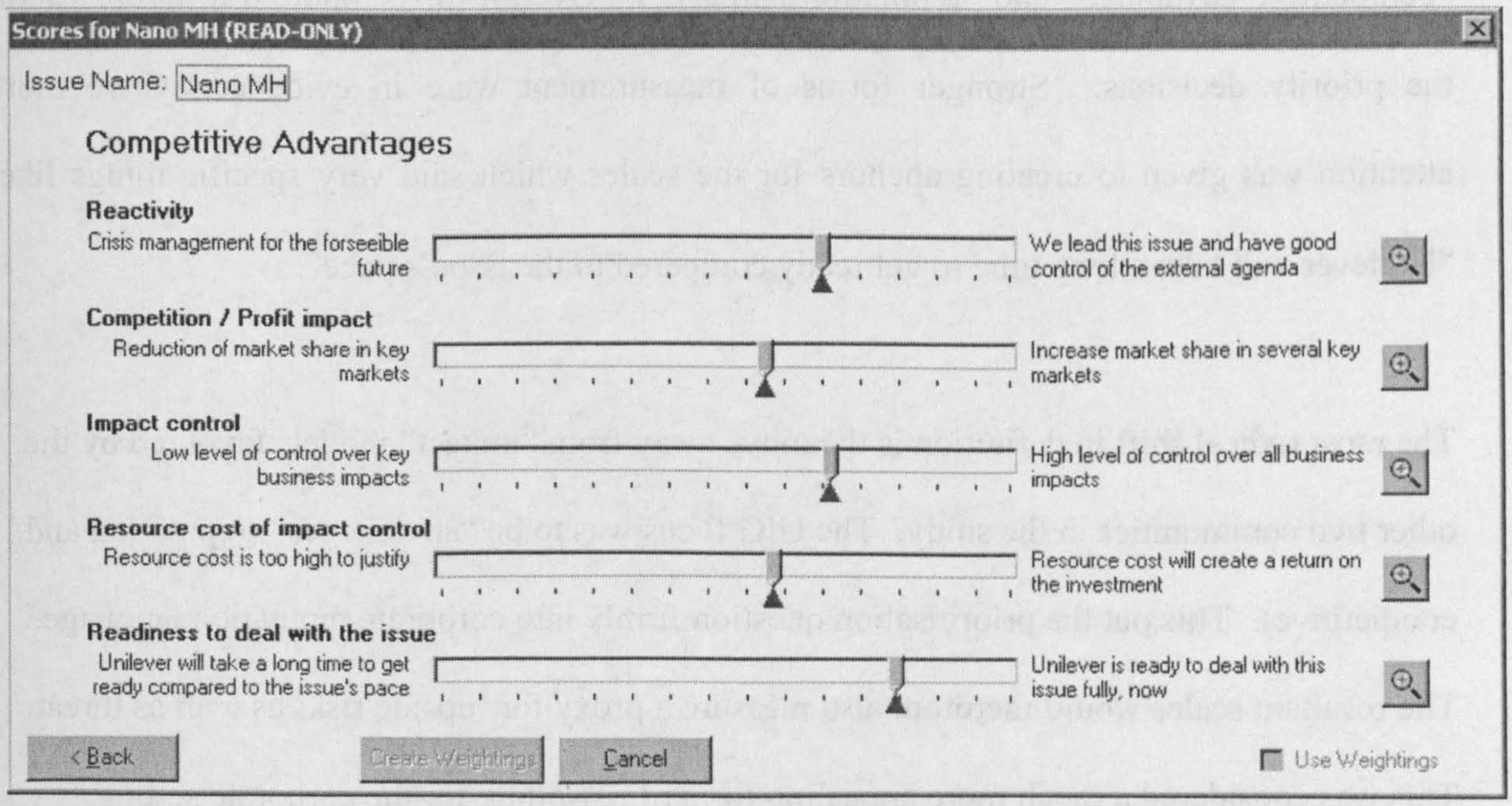
Variable	Measure or comment		
Perception	Stakeholder analysis or current IP metrics in tool	Ascertaining in some useful way how the issue is being perceived and what the intentions are of other key players	Characterisation and classification
Could use existing perception attributes: Intrinsic value, visibility, ownership and business impact,	Categorical slider bars	correlation with subjects of ngo stuff, linked to impact on corporate reputation	

The final score-card

Many iterations were required to produce a score-card which could satisfy only some of these aspirations. When this was finally agreed the threat and impact language of previous score-cards had been partially replaced by more positive "reputation and competitive advantages". Further improvements in the detail of behaviourally anchored rating scales were also seen.

Figure 39: The first UIG score-card





Review of this score-card

The attributes are beginning to look more like elements of a formal scale here which is tapping into an overall construct. The anchors are highly meaningful. Each scale has its extremes defined in a realistic behaviour which is essentially measurable. Reading vertically down the anchors of the right hand side of these scales encapsulates a "grand heuristic" to assess how well issues are managed. It is a short step from there to give those issues some

kind of data-driven priority estimate. What is also very interesting is that the "grand heuristic" for a well or poorly managed issue in Unilever's eyes is visible as a narrative in the score profiles.

Improvements brought about by this case

This case demonstrated that global issues prioritisation on a common scorecard was not only possible but that it could create the driver for new reasoning heuristics to come along which the group would accept and act upon. The benefits created in the earlier cases were still seen to be transferred here: systematic combination of multiple scorers views; use of multi-attribute scales; emphasis on behavioural anchoring; encapsulating key metaphors (like "competitive advantage" and "reputation damage"). Overall this supported a rational audit the priority decisions. Stronger forms of measurement were in evidence, where more attention was given to creating anchors for the scales which said very specific things like: "Unilever will take a long time to get ready compared to the issue's pace".

The most radical shift in definition is the move away from "impact" models favoured by the other two communities in the study. The UIG focus was to be "advantages" (reputation and competitive). This put the prioritisation question firmly into corporate reputation language. The resultant scales would therefore also measure a proxy for 'upside risk' as well as threat. This was considered a much more appealing form of reasoning for the corporate setting.

Business impact was however, retained as this was for a crucial technology transfer argument as it was obvious from the previous cases that it would remain important to talk about impact. Communities who 'traded' in impact assessments were sceptical of the up-beat language of business advantages seeing it as "corporate spin". A benefit of behaviourally anchored rating scales is the ability to hold the two views in tension in the same measurement.

Increased process rigour: The prioritisation process was designed with three elements:

1. The global issues team would (independently or in plenary) perform an analysis of their issue using the new score-card.

2. The global issue leader would be responsible for ratifying the team's scores to one view.

3. The UIG would be presented with the results in order to decide on the priorities.

Improved heuristics: In all three cases the 'essential natural heuristics' which people were operating with to create prioritisation (defining, scoring, analysing and communicating) had all been demonstrably improved by their score-card developments. These improvements came about not by introducing a model which supplanted existing preferences, but by introducing a model that mirrored and improved them. This opened an less threatening door to change existing heuristics to ones which were more formal and analytical but still recognisable.

4.3.4 A strategic impact score-card

The UIG score-card had, out of the three on offer, the strongest link to "corporate business drivers". This was achieved through the emphasis of advantages over impacts. This led to a further research challenge. This was the communicate the prioritisation results to senior stakeholders fully in the language of their own business drivers. The answer to that question created a powerful step change in reasoning about priority.

A step-change in reasoning

In the first round of global issues prioritisation nearly 80 issues were taken through the scoring process. Under review the prioritisation tool and process were well received, the results from the global issue teams and leaders using the current score-card were considered to be a definitive "technical" assessment of the issues portfolio.

However, the absence of stronger rhetoric of "business value drivers" for senior communication was still considered a weakness. It had been expected that these strategic ramifications of the priority assessment would be a more conventional discussion process using use the expertise of the UIG themselves and that this process would lead to their translation into this sort of language. The benefits of a score-card approach were considered to be so strong that it was decided to extrapolate the research into an additional, strategic version of the entire prioritisation process. This meant the development of a new score-card to assess the strategic relevance of priority technical issues.

Strategic prioritisation

The source of the strategic criteria would be "SIA". This stood for Strategy Into Action plan. Preserving the need to put the results of this further analysis in plot form, the UIG chose impact on SIA (positive and negative) and "influence" as the variables. A workshop was held to develop the influence criteria as these were a matter of professional judgement.

For the alignment with SIA a more formal method was chosen. SIAs, by their nature, are very high level, hierarchical tools. That is to say Unilever's board sets the company SIA and then there is a cascade of further SIAs in response. Three SIA plans (the foods category, the HPC category and the SIA of Unilever board) were examined in detail.

Deconstructing Strategy Into Action

An SIA takes up a single page of a Powerpoint slide. The core structure is fixed:

- Must win battles
- Strategic goals
- Key Performance Indicators and Targets
- Strategic Actions

Stage one: Actual statements

Each of the SIA plans was broken down into the number of "statements of intent" it contained excluding explanatory text, such as the titles above.

UEx = 69

HPC = 64

Foods = 51

Removing duplicates left 116 unique statements across the three SIAs. These statements could be placed into the 39 categories shown here:

Figure 45: SIA content analysis

Global mindset	Optimisation	Health
External orientation	Effectiveness	Consumer insight
Alignment	Responsiveness	Customer insight
Competitive		Channel strategies
	Portfolio choices	Rennovation
Key Markets	Regional / global mixes and category strategy	Vie, Soy, Lipton, HH emphasis
D&E Markets	Financial flexibility	Distinctiveness development
Mature markets	Confidence in leadership	Functional through relevant science
Stop decay	Real accountability	Sharper positioning
Accelerate growth in china and russia	Complexity reduction	Effective brand communication
		Inspiring activation platforms
Deliver top ten	Marketing	Vitality partnerships
Link customer insight	Customer development	Vitality incorporation in brankey
Link consumer insight	Information mangement	
	Communication	
	Leadership	

Stage two: Clustering

The 39 categories clustered into 7 key areas shown below

Figure 46: SIA clusters

Unilever Outlook	Global mindset	**Markets**	Key Markets
	External orientation		D&E Markets
	Alignment		Mature markets
	Competitive		Stop decay
			Accelerate growth in china and russia
Organisation	Regional / global mixes and category strategy		
	Portfolio choices	**Brands**	Health
	Financial flexibility		Consumer insight
	Confidence in leadership		Customer insight
	Real accountability		Channel strategies
	Complexity reduction		Rennovation
			Vie, Soy, Lipton, HH emphasis
Innovation	Deliver top ten		Distinctiveness development
	Link customer insight		Functional through relevant science
	Link consumer insight		Sharper positioning
			Effective brand communication
Supply chain	Optimisation		Inspiring activation platforms
	Effectiveness		Vitality partnerships
	Responsiveness		Vitality incorporation in brankey
Capability development	Marketing		
	Customer development		
	Information mangement		
	Communication		
	Leadership		

Stage three: Identification of common measurement

Each of these 7 areas had two or three main "impact variables" identifiable within it:

Figure 47: SIA main areas

High Level	Main variables
Unilever Outlook	Global-Unilever, externally competitive
Markets	Share, growth, recovery
Innovation	Focus, success, insight
Supply chain	Optimise, effective, responsive
Organisation	Strategic, simpifcation, leadership
Brands	Insight, communication, health/vitality
Capability development	Leading, understand, communicate

The final scorecard

Armed with the above analysis and the variables for "influence" the final score-card was worked up through drafting and testing. As with the previous score-cards in this study the aim is to develop a robust set of questions which could be rolled up into a single metric.

The final strategic score-card took the following form

Table 17: The SIA score-card

SIA "Fit" scorecard		
Supply chain: What overall impact will this issue have on either the optimisation, the effectiveness or the responsiveness of Unilever's supply chain		
Very negative impact on Supply chain	No real positive or negative impact on Supply chain	Very positive impact on Supply chain
Vitality: To what degree does the management of this issue bring significant delivery on nutrition, health / hygiene and vitality		
Left un-checked it damages delivery	Some contribution with links to thought-leadership	Contributes heavily to thought leadership (with key external audiences)
Innovation: To what degree can the management of this issue help with the delivery of the "top ten" global innovation projects, in either category, on time and in full		
Left unchecked likely to hinder top category innovations	Some generic contribution but uncertain to quantify	A direct and defined link enabling delivery on one or more top category innovations
Distinctive brands: To what degree can the management of this issue bring significant contribution to unique, insightful, functional, competitive, scientifically proven brand benefits		
There is a no argument that it will help and may damage key brands	There is a reasonable basis to assume a reasonable contribution	There is a compelling argument for a powerful contribution
Growth: To what degree can the management of this issue bring significant delivery on "the creation of great brands preferred by retailers and repeatedly purchased by consumers"		

Left unchecked likely to lead to downturn in sales / not achieving growth targets	Some contribution to growth, but hard to quantify	A compelling case for a direct contribution to brand/market growth performance
Key market strategic actions: Does the management of this issue contribute to Unilever's "strategic portfolio choices" as expressed in the SIA "win key markets"		
Damages delivery on either key Brands or strategic market targets	This is linked to Unilever's strategic portfolio choices	There is a direct and defined link to key brands and strategic market targets
Winning with customers: To what degree can the management of this issue bring significant contribution to how we meet our "win with customers" aims (as specified in the SIA)		
Negatively impacts on our win with customers SIA aims	Some improvement will be possible but it is difficult to quantify	Positive impact creating competitive advantages for Unilever and customers

Table 18: The influence and control score-card

Influence and control scorecard		
Unilever ability to influence :To what degree can Unilever have significant influence over the way this issue is discussed and the way it plays out :		
Very low level of influence	Have influence but not able to control	Very high level of influence can act as a leader
Unilever's need to have an influence : To what degree **should** Unilever have a significant influence over the way this issue is discussed and the way it plays out :		
This is not an area for Unilever "ownership" this is a sector issue	Unilever needs to be a significant force in the debate and this is expected	It is crucial and expected that Unilever will lead and drive this issue
How does Unilever need to drive or lead on this issue:		
Low external visibility, dealt with in the context of a 'sectoral' debate	Some external visibility but in coalition with other companies and relevant stakeholders	High external **leadership** visibility because the issue is crucial to Unilever's competitive priorities
Impact shift: If the **impacts** of this issue came to our door to what degree can we control the outcomes to our competitive advantage		
Very low level of control with likely disproportionate damage to Unilever (compared with competitors)	Some control to share the impacts across Unilever and competitors	Very high level of control, allowing competitive advantages by significantly shifting negative impacts (to competitors)

Important to note about this score-card is that the propositions have become more detailed. Also, because these propositions are quite a bit more concrete and complex than those used before, a central behaviour anchor was introduced to stabilise the meaning of the rating scale.

Improvements brought about by this case

There were some interesting improvements in this case in three areas:

- in the meaning of what the heuristics are trying to do
- in the process by which this approach and tool are deployed
- in the way the results are visualised

Improved scales: The meaning of the scales being used in the global issues teams' version of prioritisation was already based on a solid set of questions with behaviourally verifiable anchor points. The opportunity afforded by access to the strategic documents which set Unilever's business priorities however, opened up a new opportunity. The deconstruction of the strategy into action plans gave the most grounded score-card yet. This is evidenced by the more proposition like nature of the questions and their increased specificity. Three scale anchor points defined the behavioural grounding of the answers more firmly.

Improved process: The process by which prioritisation takes place had reached a fully mature stage spanning the gap between technical project areas and strategic priority setting. There were now two scorecards. One was to rationalise a kind of solid and sensible operational priority, facing broad drivers like competition and reputation. The second was used by a more elite group was to rationalise the "meaning to Unilever's strategy". The result was five stage prioritisation process. Each stage maintained the same rationality and was bounded by similar data.

Improved argument: Lastly, in terms of improvements, although the function to compare users at the score-card level had been available the whole time, the strategic assessment was the first place where it was formally used by the (highly divergent) group of decision makers to resolve key disagreements.

4.4 Unilever case study two: The development of visualisation environments for issues prioritisation

Introduction

Three projects to deliver four score-cards to user groups within the Unilever issues management community have been described in the previous section. These projects ran essentially in a sequence. Each of these score-cards, as well as being a visualisation itself, produced results which the users desired to view in a range of Cartesian plots. In a parallel exercise with score-card development, these plots were also evaluated and improved. This case study considers how this was done and the implications for the contribution to support for heuristic reasoning which the plots delivered.

4.4.1 Aims

This case study describes the applied development of three issues prioritisation visualisations, that work can be summarised under the following aims

1. To work with subject matter experts to extract user requirements
2. To design and evaluate a series of prototype visualisations.
3. To evaluate final versions in user trials

4.4.2 Objectives

The part of the projects to design these plots can be described by the following (sometimes overlapping) objectives:

1. To assess existing practices in issue scores plotting
2. To develop the user requirements for improvements in the light of new score-cards.
3. To derive separate, business unit specific plotting devices.
4. Internal testing and validation of plots.
5. Iteration of plots design.

6. External testing, validation and final improvements to plots.

4.4.3 Methodology

The following methods were used (to differing degrees) in development of the visualisations.

- Observation of existing practices in issue prioritisation visualisation through attendance at project meetings and discussion with project users.
- Structured group discussions with commissioners to assess the suitability of business unit specific visualisation content.
- Structured and ad-hoc one to one interviews with users to evaluate visualisation elements and assess these for consistency and comprehensibility.
- Plenary discussion sessions with user groups to test visualisations on prioritisation exercise results.
- Final field testing of visualisation methods.

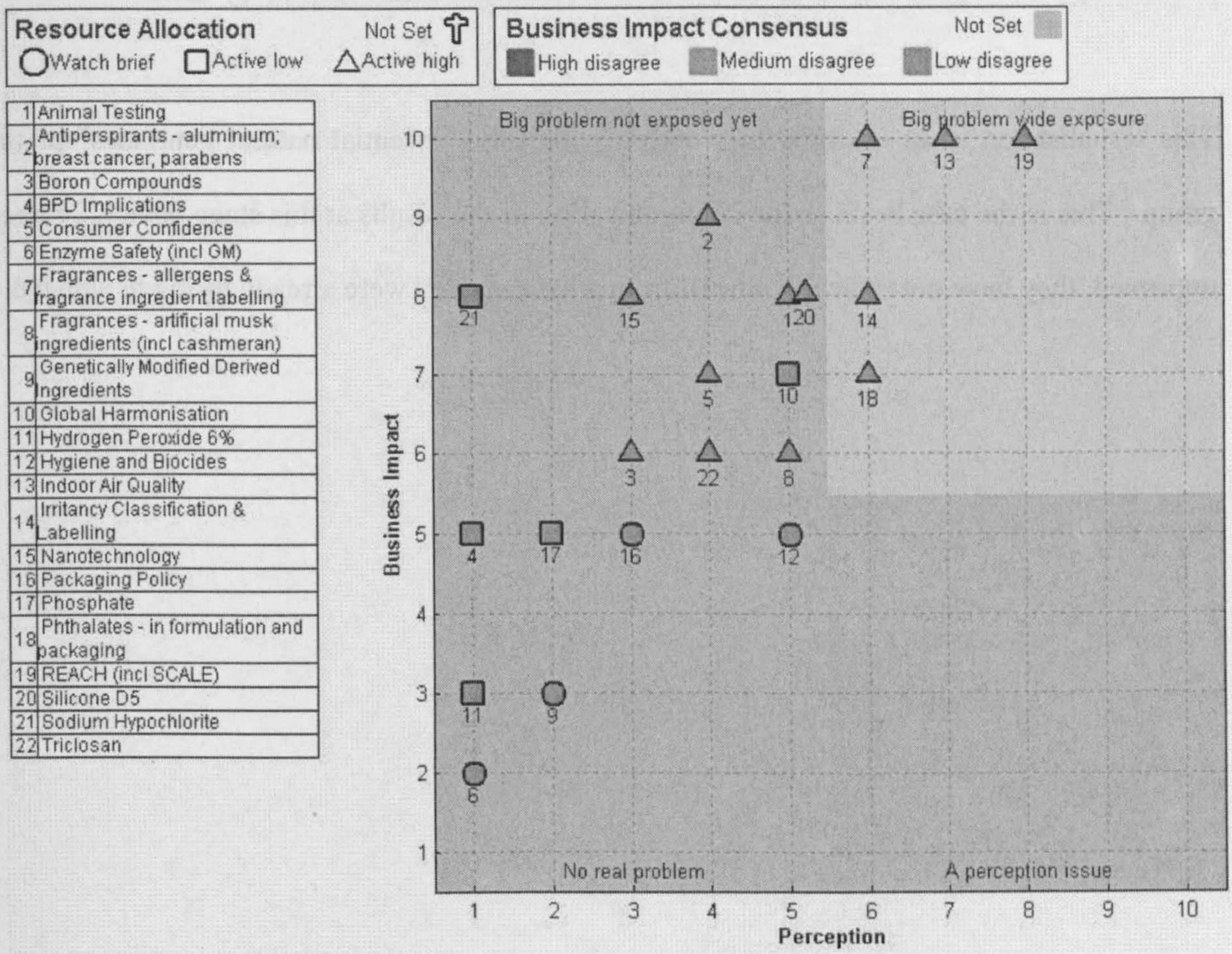
4.5 Consideration of results

4.5.1 A basic case in Unilever's HPCE category

The final results of the HPCE group's prioritisation were published in the form of Cartesian plot split into equal quadrants with data points shown as small shapes. Improvements on the score-card side and in the scales of measurement made it possible to improve the existing plot design (which was 'drawn' in Powerpoint). Numerical scales could now be added the quadrants could be more accurately (still in the heuristic sense) be labelled in behavioural terms.

A software prototype was used to create a graphical user interaction based plot in a stand alone application. The first improvements to the HPCE approach are shown:

Figure 30: The HPCE issues prioritisation after changes



This visualisation is in fact a very sophisticated version of the flip chart drawing one might use in a post it pad exercise. It was considered important by users to retain the overall flavour of this familiar heuristic device. With the improvement in measurement scale came the ability to use numbers on the axes in a representative manner (note these were still not really a mathematical scale). The qualitative characterisation (i.e. the interpreted meaning of the data points) has been improved by the power of the software to draw and revise shapes and colours based on underlying scoring which can of course be changed at any time.

This early case of HPCE's improved visualisation was a good example of the use of a software solution and on the acceptability of modest improvements via an interactive

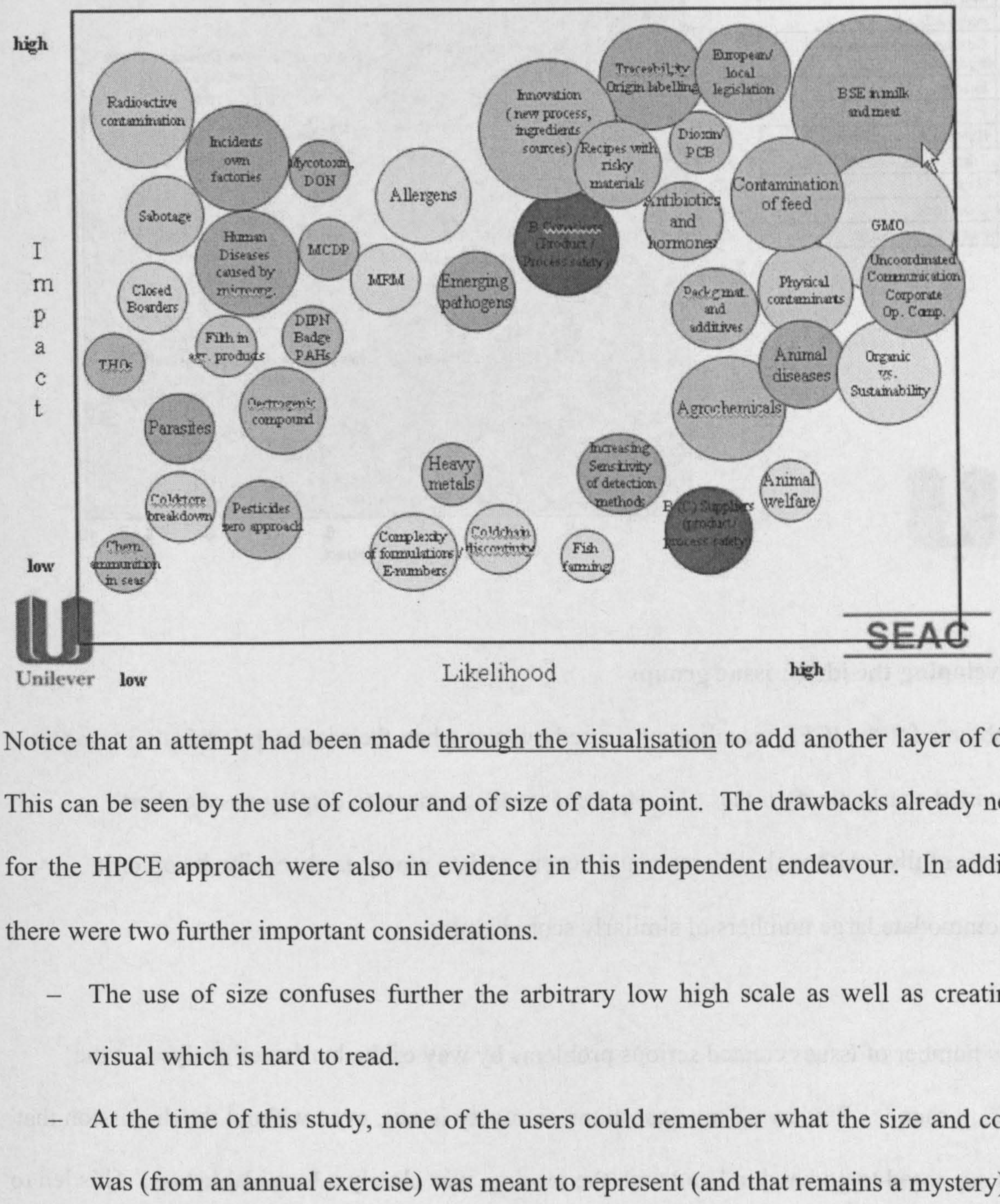
graphical tool to replace post it pads leading and the resultant "cleaner" representation of the problem space. Importantly, these improvements had also been accepted by users.

This visualisation is an example of working within the 'essential natural heuristics' of the group. This is the case because the communication of the results at this stage have only been improved, they have not radically altered from what the users were already trying to achieve.

4.5.2 **Case two ICFE visualisation needs**

Unilever's Ice-Cream and Frozen Foods category (ICFE), like their colleagues in HPCE, wanted to visualise a portfolio of issues. The results of their approach had likewise been drawn in Powerpoint. This is shown below.

Figure 32: The ICFE 'Bubble Chart'

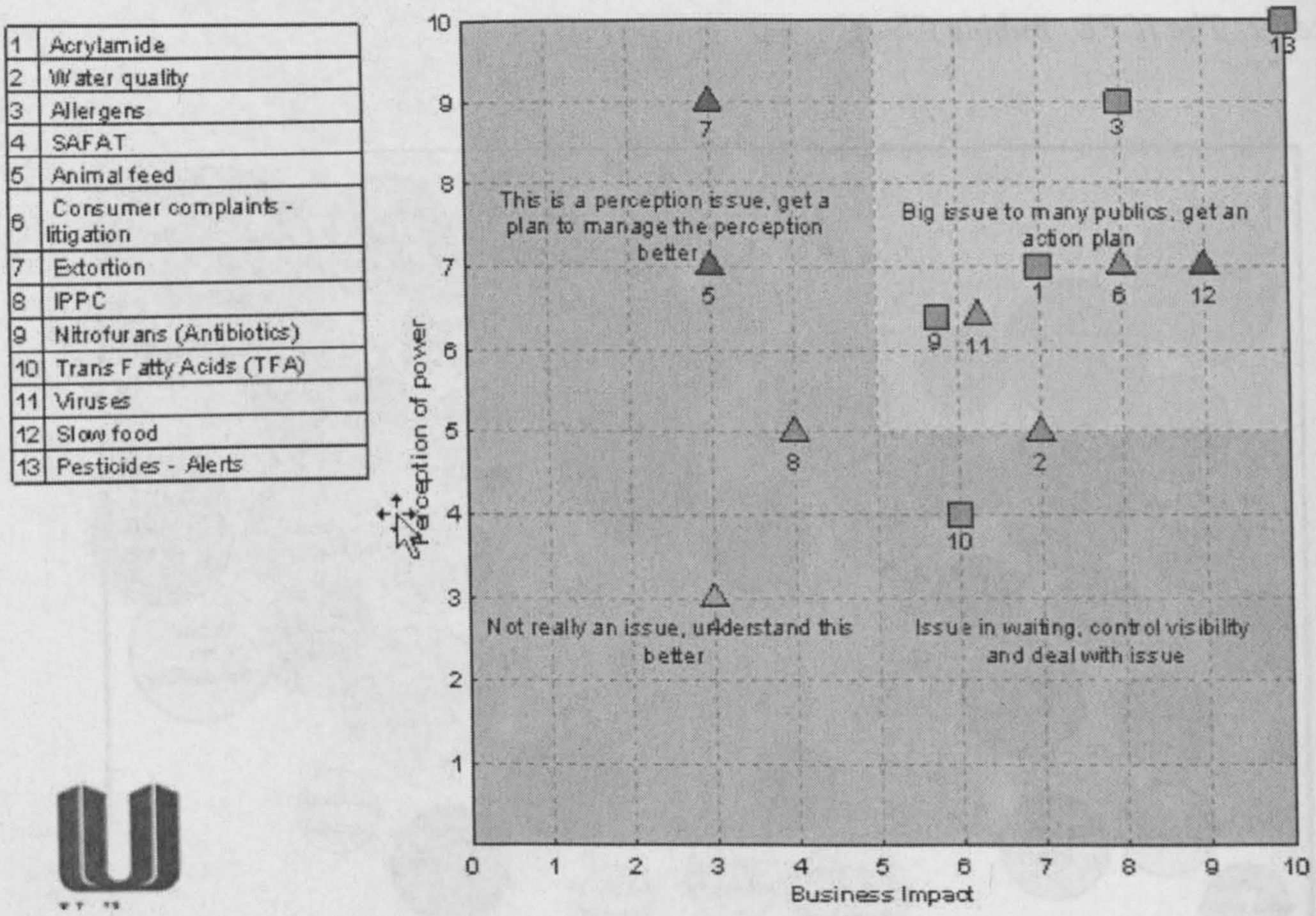


Notice that an attempt had been made through the visualisation to add another layer of data. This can be seen by the use of colour and of size of data point. The drawbacks already noted for the HPCE approach were also in evidence in this independent endeavour. In addition there were two further important considerations.

- The use of size confuses further the arbitrary low high scale as well as creating a visual which is hard to read.
- At the time of this study, none of the users could remember what the size and colour was (from an annual exercise) was meant to represent (and that remains a mystery).

The first trial of the new score-card for ICFE issues prioritisation was accompanied by an adaptation of the HPCE visualisation. This is shown below. Notice this retains the essential flavour of HPCE's version but the quadrant labelling is far more detailed.

Figure 34: ICFE early prioritisation results



Developing the ideas: issue groups

Problems for the ICFE visualisation arose however when their issues prioritisation process provided nearly ninety issues. Ninety issues was too many to display on one chart meaningfully. Although the area could accommodate ninety fairly easily, it couldn't accommodate large numbers of similarly scored issues.

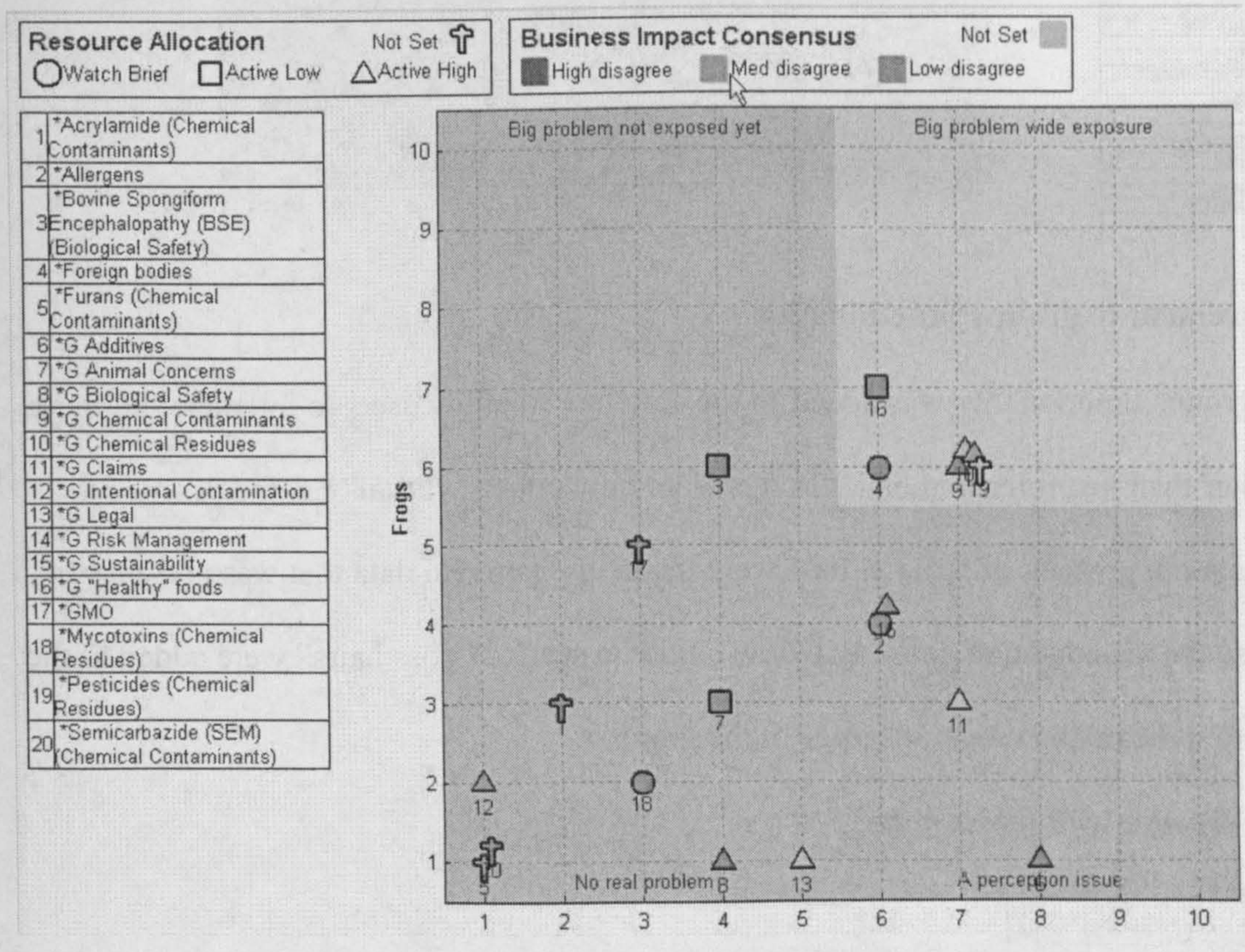
This number of issues created serious problems by way of the burden of analysis in the scoring rounds. Reviewing nine questions on ninety issues, even without the discussion that was supposed to go hand in hand with the scoring, was clearly a formidable task. This led to two innovations in the design which were both supported by variations on the visualisation.

Both of these were forms of issue grouping within the tool. One was called 'issues groups' and the other was an issues 'car park'.

Issues groups

Initially working issues into groups had to be done in a rather clumsy manual fashion. In the diagram below items prefixed with a 'G' refer a cluster of similar issues. The creation of these eleven groups made a portfolio of ninety manageable at twenty issues. The meaning of the groups was being mediated through the visualisation.

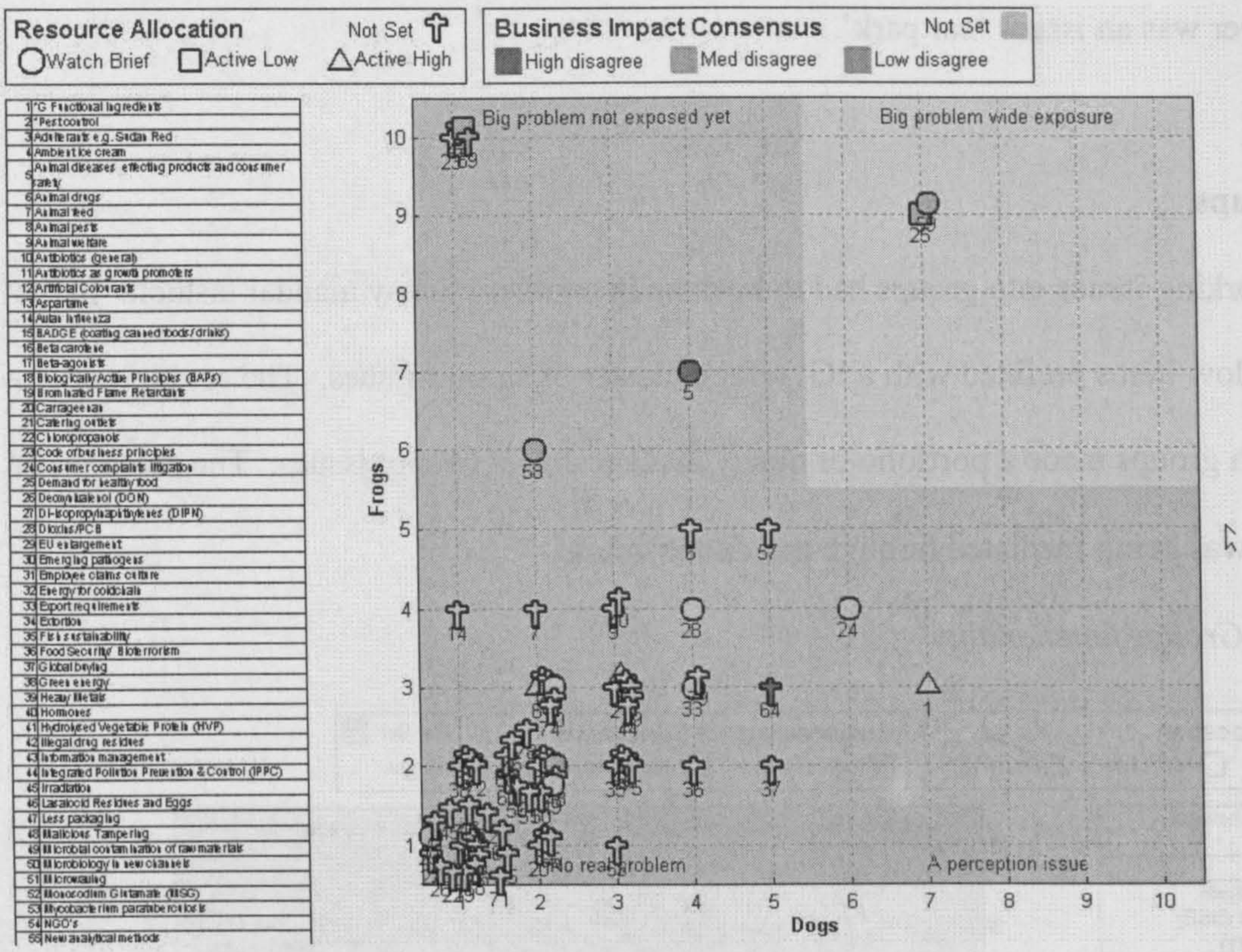
Figure 35: Groups functionality



Issues car park

A reluctance to lose the data on issues which had been scored into groups created a second visualisation. This was named, after a popular group facilitation technique, the car park. A car park is a place, normally a flip chart in group process, which is reserved for placing items which the group will not be discussing. The car park from the first ICFE prioritisation round is shown below.

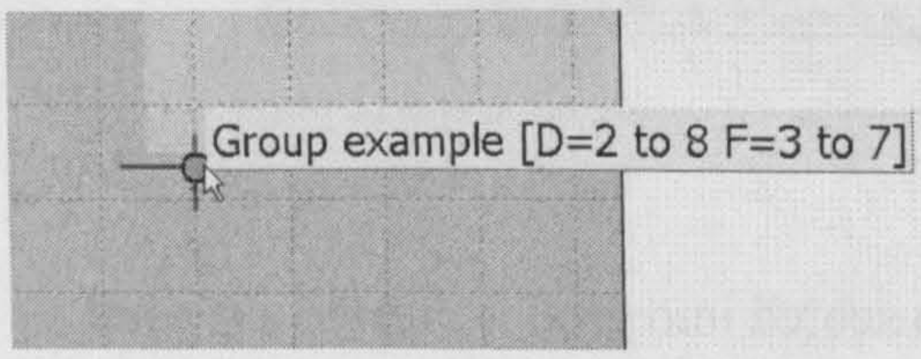
Figure 36: The issues 'car park'



An improvement to groups functionality

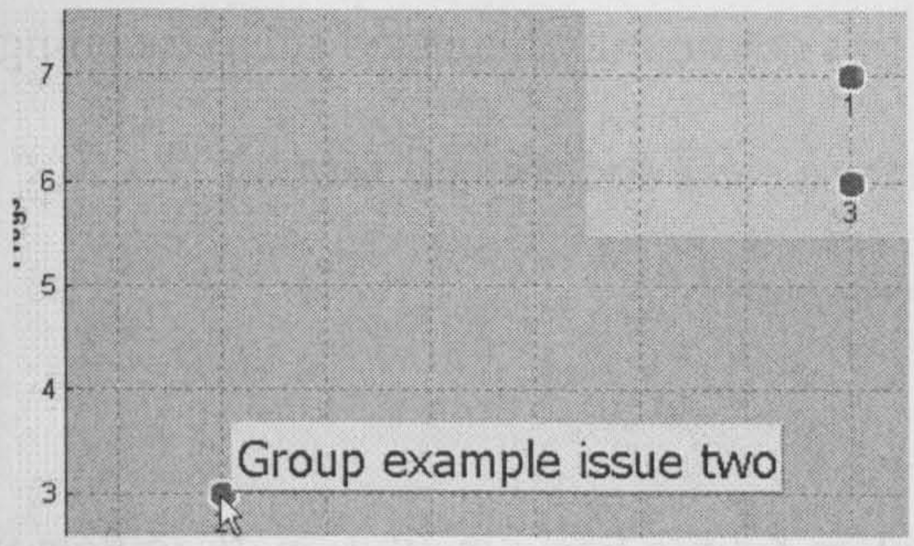
A formal groups functionality was added to the software to allow users to interact with these groups rather than just refer to them. Groups were now either "virtual" i.e. the denoted a group of data not present, or real i.e. they were made up from real data that was present. In the real case the visualisation cutaway below could be seen. "Cross hairs" were added to the data point to indicate the range of scores in the group.

Figure 37: Groups with 'cross hairs'



Further functionality also allowed one or more groups to be 'exploded' to reveal the underlying data as is shown in this cutaway.

Figure 38: Groups at next level of detail



Improvements

The sophistication of the way in which the visualisation is now supporting reasoning as well as displaying results has become very interesting here. This is an example where the need to visualise the data in a way which also manages it has led to a successful adaptation and expansion of the natural heuristics of the group.

The value of this visualisation is shown in the fact that when the software visualisation was swamped with the data, the user group, convinced by the improved process, decided to adapt their approach to this limitation. In truth it was an underlying limitation of their own process also. So both the process and the software had to adapt. The fact that they did not return to a more simplified visualisation proposition is a testimony to the benefits of the new visuals.

The introduction of grouping of issues to reduce data volume added a second tier of data representation (the group plot) . The complexity of the visualisation environment was thus increased, as was the sophistication. Now very large numbers of issues could still be accommodated and, only when needed, was their data reviewed.

The HPCE case had demonstrated that you can work within the 'essential natural heuristics' of a decision making group and generally improve them. What the ICFE case study does however is demonstrate that you can also expand the essential natural heuristics of a user

group. You can do this in a reasonably sophisticated manner as was seen in the innovation of grouping issues. This was a step change for the way this community managed such reasoning tasks. This change was primarily driven by the need for a clear and concise form of visualisation without an associated loss in data sophistication.

This group of prioritisation users were successfully taken from a fairly institutionalised Post it pad exercise, the visualisation of which they struggled to remember the meaning of, to a visually very adroit and data driven audit trail of reasoning. The group accepted the approach and the technology into their reasoning behaviours and reported benefits to that reasoning and its resultant communication.

4.5.3 Case three: UIG visualisation needs

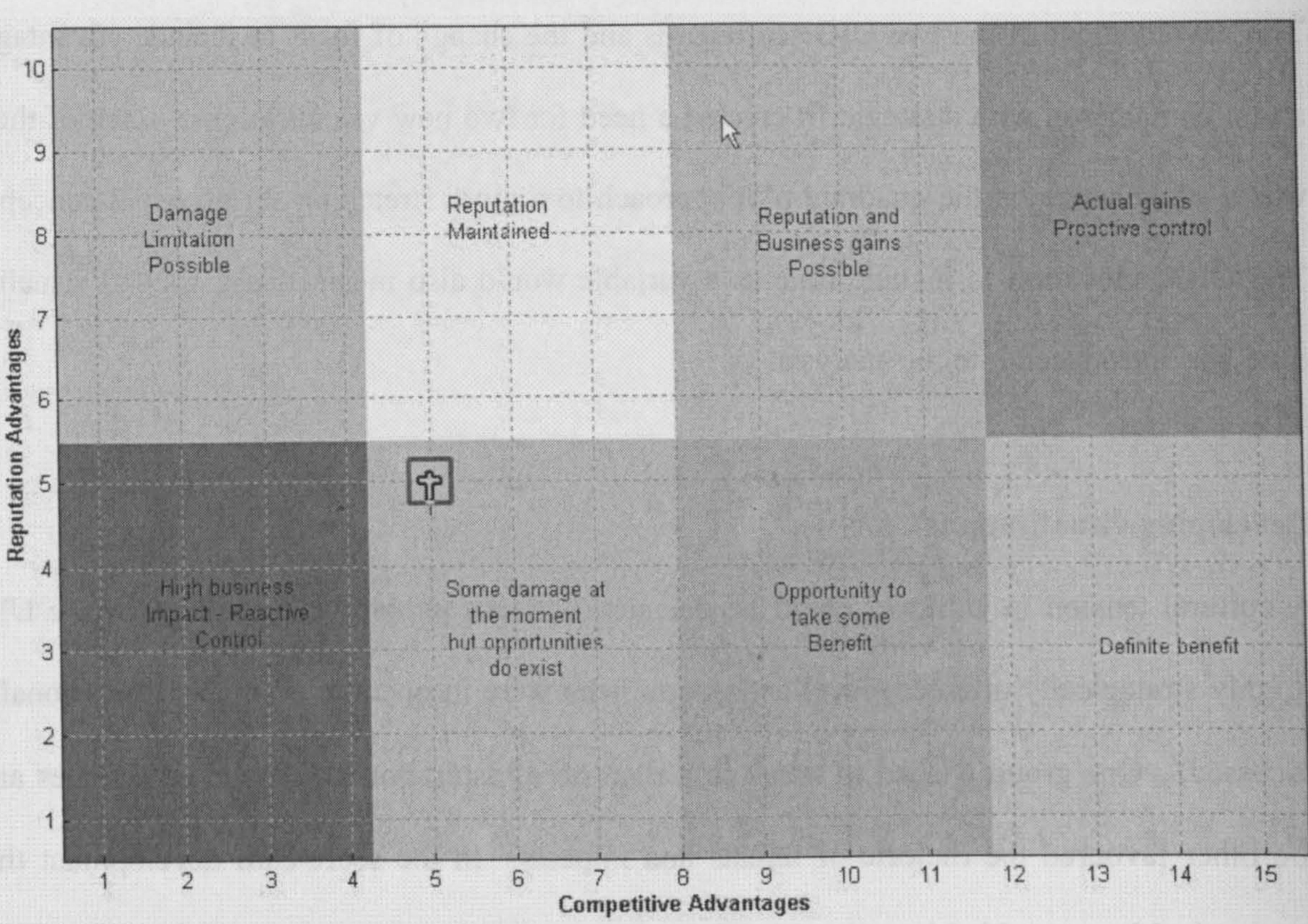
The development of the two UIG score-cards and the change of focus to include advantages and a comparison with strategic fit created a need for two new visualisations. Each of these was a step change on the quadrant plot approach to a more stretching set of visual concepts. The UIG innovation to include time as a variable would also meant that a way to visualise time was introduced into the analysis.

Developing visualisations

A cultural tension in Unilever could be detected between people who operated at the UIG (highly strategically oriented) level and people who were in operations (highly operationally focussed). One group wished to speak in a rhetoric of aspiration all around advantages and the other favoured the rhetoric of threats and impacts. In the score-card development this tension had been avoided by using bi-polar scales for scoring. In the visualisation however, three further innovations were needed and these would increase the visual complexity of the heuristics being used.

Contradictory needs to focus on impact and advantages were resolved by effectively gluing together two plots (one threats and one advantages) in a single visualisation. The result was a more continuous plot whose aspect ratio stretched one of the scales. The plot areas were carefully labelled to reflect the duality of the users' worlds.

Figure 40: The octant plot



This increase in visual complexity did cause some concern it is important to note that these tools were being developed in a world very fixated on simplification. Injecting new ideas and the visual complexity one sees here is very difficult.

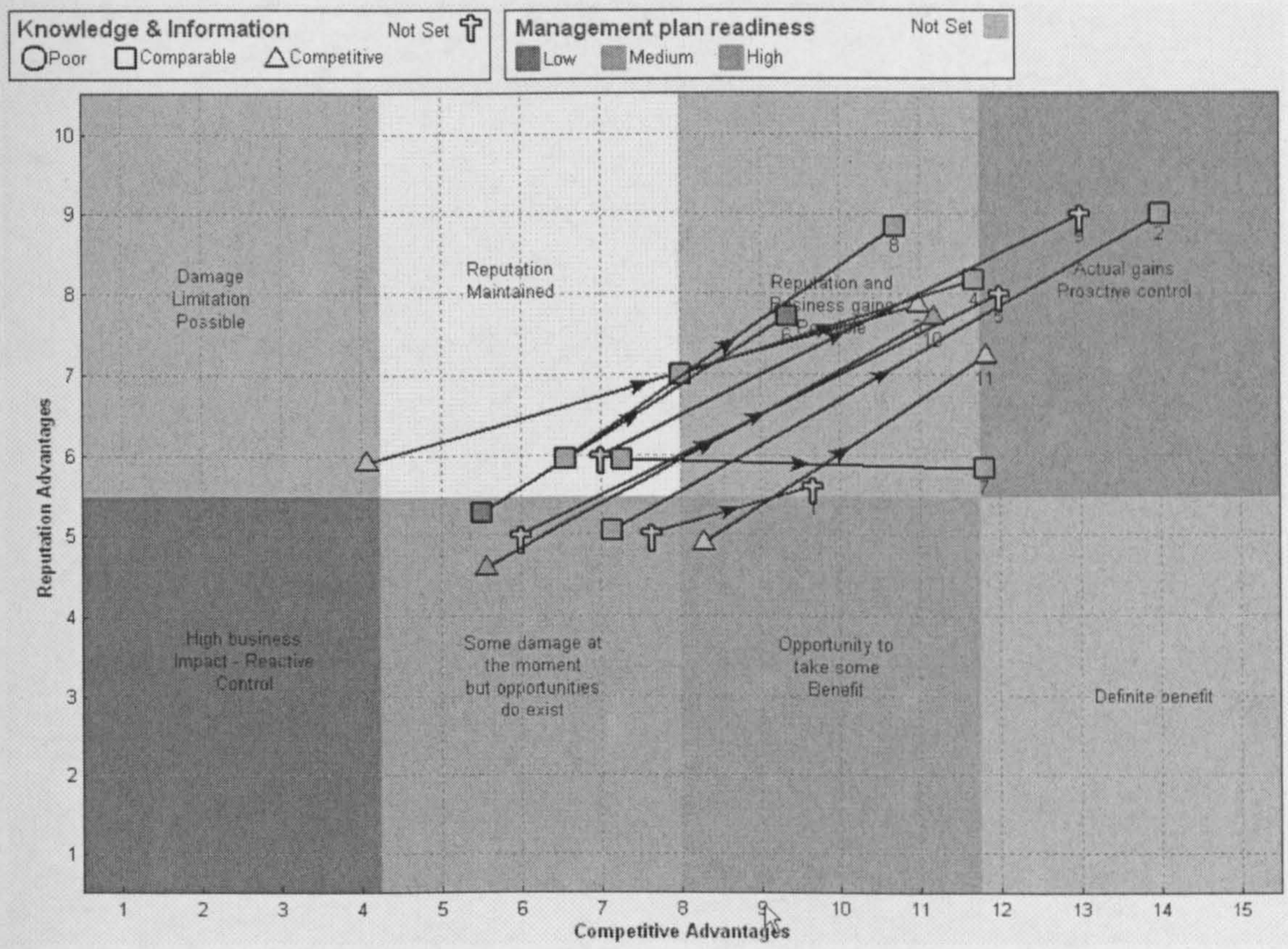
A second innovation however would see the plot widely accepted. This innovation would also result in a sea change in the way the company thought about issues. In Hurisk it had always seemed sensible to add another dimension to risk by thinking about it now and in the future. That idea was even more effective in Unilever.

Time and change

"Please rate your issue as it will stand if managed really well in the future". Although there would be caveats about "realism and grounded thinking", this little statement significantly changed users' ability to talk about issues. The proposition was simple, the same score-card could score an issue in two time frames.

The visualisation software was modified to read two sets of data and create a new visualisation. Notice that the perceived benefit in this analysis was so great that no-one questioned the need to effectively double the number of questions answered. The impact on the power of the system to improve reasoning was noted to be very great. The impact on the use of the visualisations was similarly remarkable.

Figure 41: Adding delta values



Innovating visualisation, a boundary plot

The UIG supervised the technical assessment of its issues portfolio which would be displayed on this octant plot. Due to their desire to also prioritise issues strategically, a plot to visualise this very different score-card was needed. The opportunity afforded by the need for a further visualisation lead to the creation of a specialist form of chart intended to:

- Move away from "quadrants" and their associated problems
- Make the most intuitive use of the more detailed scales
- Be easily interpreted by people who had not provided any scoring

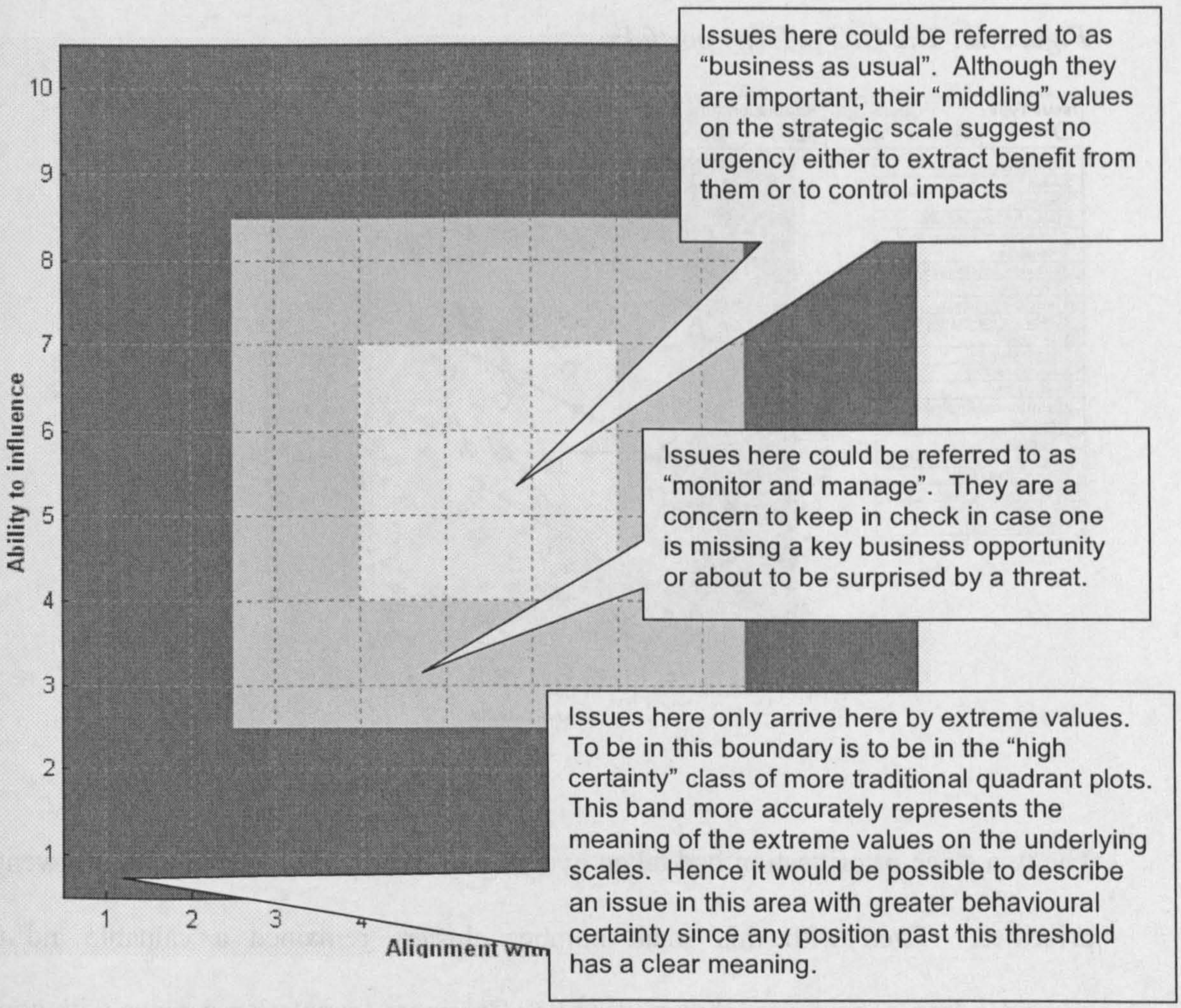
The result of this was the boundary plot shown below. The best way to explain how it is read is to use the notes provided along with the first report of strategic prioritisation presented to the UIG:

Reading the strategic graph

The strategic graph has two key differences to other similar graphics used in Unilever

a. The scales used in the scorecard are "behaviourally anchored ratings scales" using a mid point method and designed to produce interval level data. This means that the numerical position on the graph relates to a specific behavioural outcome rather than "agreement" on a more blunt bi-polar rating such as "high- med - low".
b. The graph is **not read,** in the first instance, in four categorical quarters. The use of an anchored scale with a mid point allows for the graph to stipulate meaningful "boundaries", resulting in three, as opposed to four, categories. This has the further advantage that points on the graph can approach a meaning threshold in any direction.

The figure below demonstrates the meaning of the boundaries and how to read an issues boundary plot.

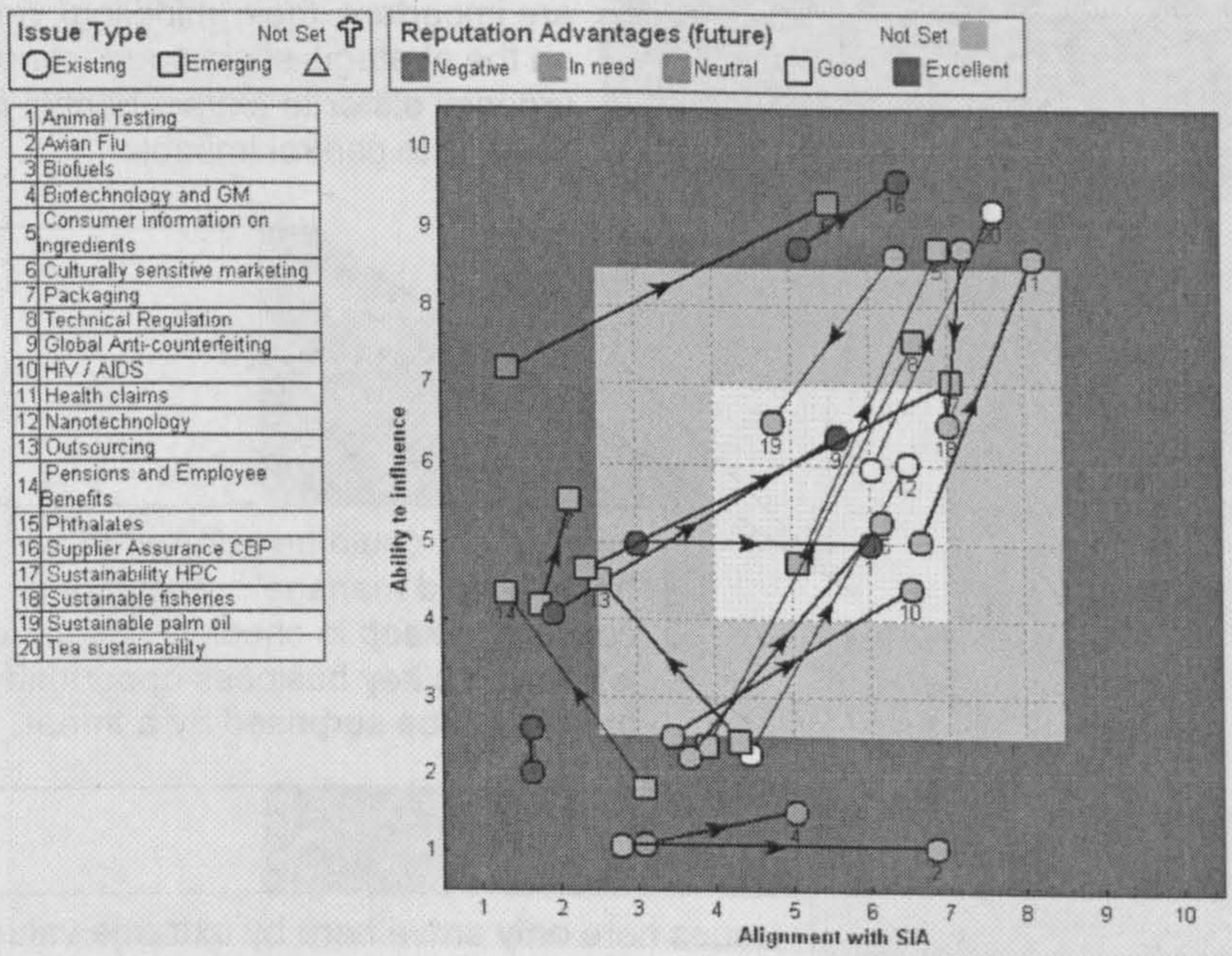


Keeping time and clustering

This boundary plot was accepted by the UIG. The techniques which had worked well in the technical prioritisation were also used here, hence analysing the issue now and in the future was retained as was the use of clusters to understand the results.

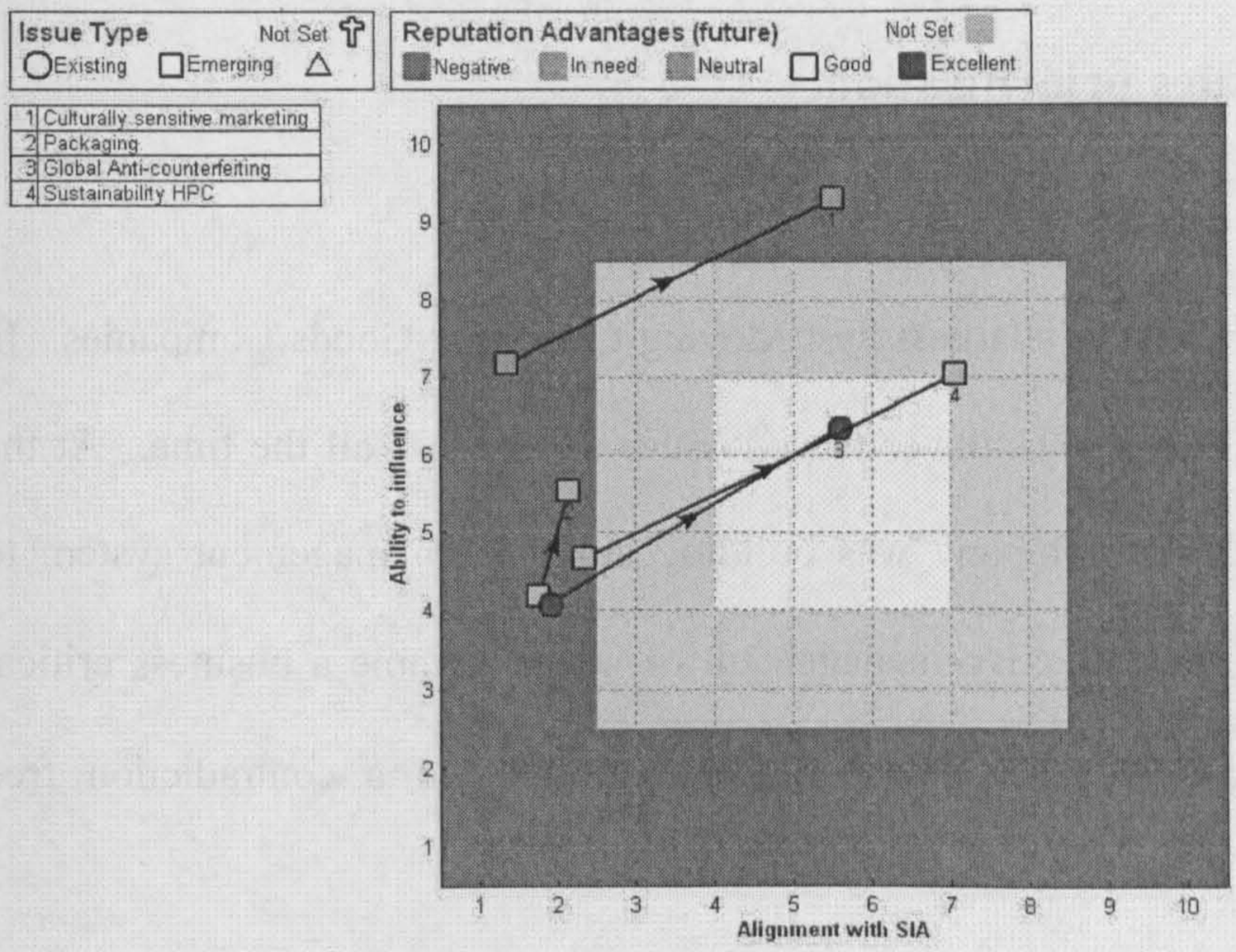
From the original technical prioritisation forty six issues were highlighted as possible priorities. Twenty eventually made it through to a final priority portfolio and their strategic prioritisation can be seen here.

Figure 48: The UIG priority portfolio



The two stage prioritisation had taken over eighty issues and reduced them to twenty priorities. Even with this small number clusters remained a valuable aid to understanding. The cluster shown was titled "business / reputation damage with good Unilever control".

Figure 49: Cluster example



Three other clusters were specified, these were:

- Business damage limitation
- Business / reputation damage with poor Unilever influence
- Business / reputation advantages with good Unilever influence

4.6 Unilever case study three: A review of heuristic concepts developed for issues prioritisation.

Introduction

Unilever is one of the world's largest Fast Moving Consumer Goods Companies. In the execution of its business Unilever faces "issues and crises" all the time. At the time of this research the company was building an issues management system to ensure that a timely and effective management of issues became a business critical activity. This would preserve continuity of business and a contradiction free corporate reputation.

The application of Descartes has been a success and its use has been extended. This system is now in operational use around the world in local and global issue prioritisation applications in Unilever. The results of two rounds of global issues prioritisation have been communicated to Unilever's board using the Descartes system. Together these would seem to suggest a high degree of successful design and transfer of the concepts and the technology.

To achieve prioritisation, effective ways for a cross section of decision makers to manage a shared understanding of Unilever's priorities was needed. As these priorities would inform significant resource decisions any tool to support this would provide a new form of decision support for risk management and mitigation.

The proposed system drew directly upon some of the concepts developed in the Hurisk project to create and test a new reasoning tool named "Descartes". Descartes was centred on heuristic reasoning techniques in a computer-based decision support prototype to:

- Support the prioritisation reasoning of Unilever's global issues leaders community
- Use a common, multi-criteria, score-card to measure relative priority of issues (even if categorically different issues are being compared, the 'apples and oranges' paradox)
- Allow users to reason with their data through a visualisation system similar visualisations in common use (but with superior utility)
- Provide evidence of an audit trail of priority issues and justify top priorities

The score-cards and the visualisations of Descartes have already been described in sections 4.1 to 4.3. This discussion paper will now review some of the wider heuristic concepts in the approach and tool.

4.6.1 Aims

The aim of this discussion paper is to highlight and comment on the main heuristic concepts developed for the Descartes tool.

4.6.2 Objectives

This paper will:

- Explain three forms of heuristic available in the case studies
- Plot ways in which improvements, modest and significant were made possible
- Demonstrate that improved heuristics gave rise to significant innovations
- Show how successful issues prioritisation process was built on these principles

4.6.3 Methods

The methods used to arrive at the data for this discussion clearly include all of those listed in the other case studies. Key methods informing the thoughts in this paper are:

- Semi-structured depth interviews with all users
- Group discussions at project meetings

- **End-user testing:** This scorecard and the visualisations were tested on groups of end users not involved in their initial design to assess effectiveness and transferability.
- **Deployment rounds:** Two Unilever issues prioritisation exercises were conducted using the tool and process. Each was closely facilitated to assess Descartes "as designed".

4.7 Discussion of results (heuristic concept development)

This discussion will surround the notion of a heuristic as described in chapter one. A heuristic is the idea of being able to reason in the absence of a complete algorithm. Three forms of heuristics will be identified and their development in the Descartes project discussed. The concept of priority

- The measurements used to create a priority estimate
- The graphs used to display the results

It will be argued that each of these objects functions in a heuristic way and that improvements to them brought about by this research also do so.

Heuristic one: Priority

Before Descartes arrived prioritisation in Unilever was primarily done using post it pad exercises. These were a very robust processes and it follows that decisions in these area were very basic also. The 'priority' they produced came from an analysis process which was not consistent issue to issue. Different issues were prioritised for different and sometimes incommensurable reasons. Priority itself therefore is the first heuristic.

Heuristic two: Measures

Closely related to priority decisions were assessments made of "probability" and of "business impact". These were clearly shown to be an impressionistic uses of these scientific terms, as the measurement scales were little more than a judgement of rank. Where aggregation was being used to reduce group scores on these measures to a single value it is clear that no calibration efforts were made. At each turn then, the measurement activity was very heuristic.

Heuristic three: Graphs

The visualisation of the results of the prioritisation was a heuristic interpretation of a scientific artefact. The results were displayed in a Cartesian graph. This graph was behaving extremely flexibly whilst maintaining an appearance of something rather more strict. The final result that was placed onto a mathematical graph was not showing the results of any mathematical calculation. Inside some versions of this graph was also placed a reference value for each quadrant which was qualitative and categorical, mixing the metaphors yet further.

Improving the measurement heuristics

Early improvements to the definition of the measurement objects, the scales used, and simple statistical treatment of data and plotting of the results were all referred to in the previous section as part of the HPCE case study. In some cases these replaced one heuristic with another. For example the categorical (high, medium, low) assessment of issues changed to a continuous scale (1 – 1000), but this was evaluated on sliding scales in the absence of any numbers.

These existing heuristics (scoring approach, the analysis and the communication of the results) were all in essence only improved, they were not radically altered from what the users were trying to achieve. These examples demonstrate that it is possible to work within the 'essential natural heuristics' of a decision making community and make modest improvements to them.

The more detailed work described earlier shows how the scales being used in the ICFE prioritisation case became multi-attribute. The variables the attributes were rolled into (business impact, and perception) clearly act as a more powerful heuristic

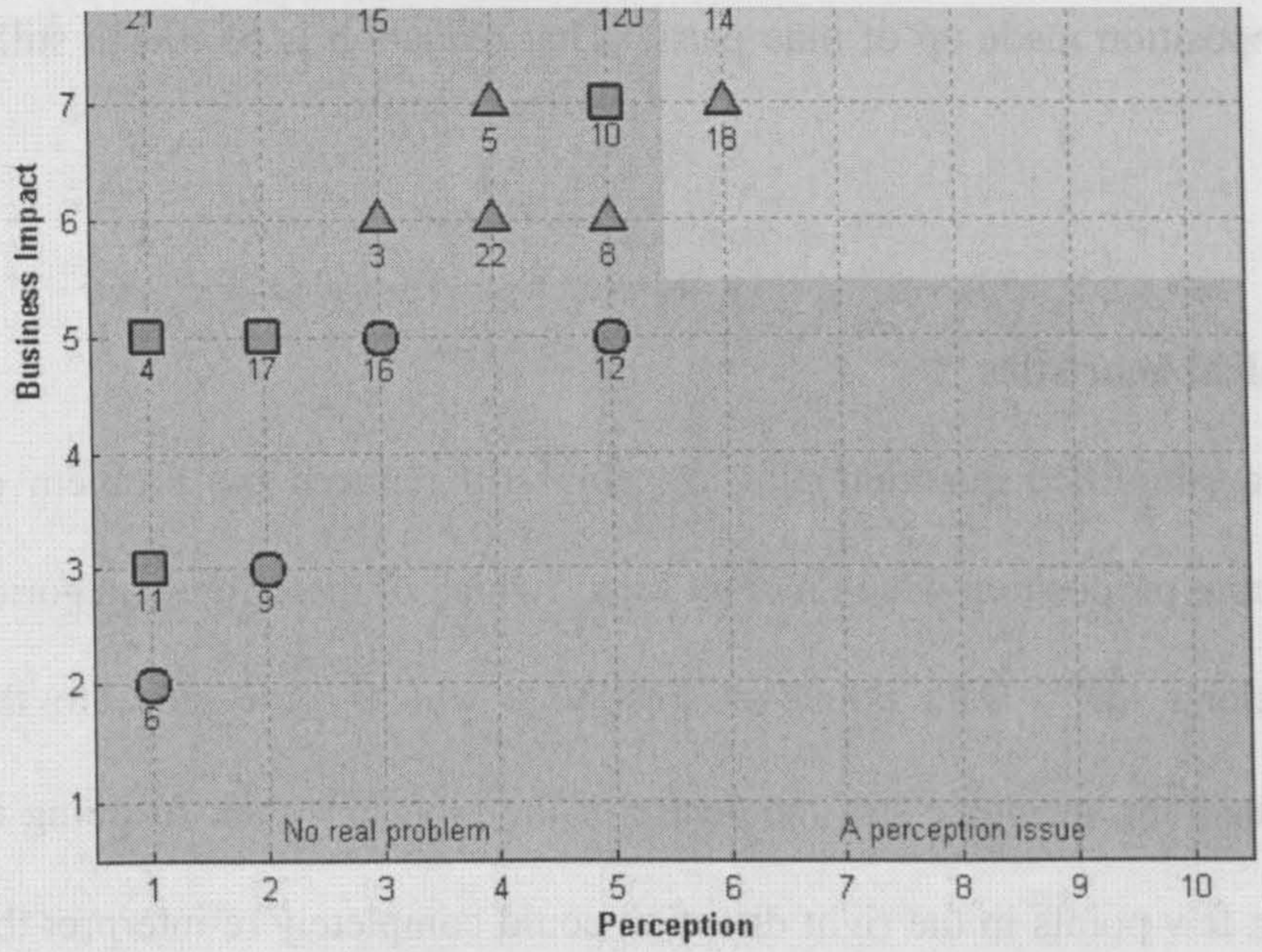
reduction of a proposition made up of nine parts. That reduction is of course still a heuristic number.

Improving graphical heuristics

The retention of a simplified quadrant plot arguably still reduced the problem of prioritisation to a core proposition which merely says: "which of these four categories does the issue belong to?" Data points in quadrants which come close to the boundary of the quadrant, or worse still the centre of the plot, lose their meaning in such a scheme as a few points in the right direction could completely re-interpret the data.

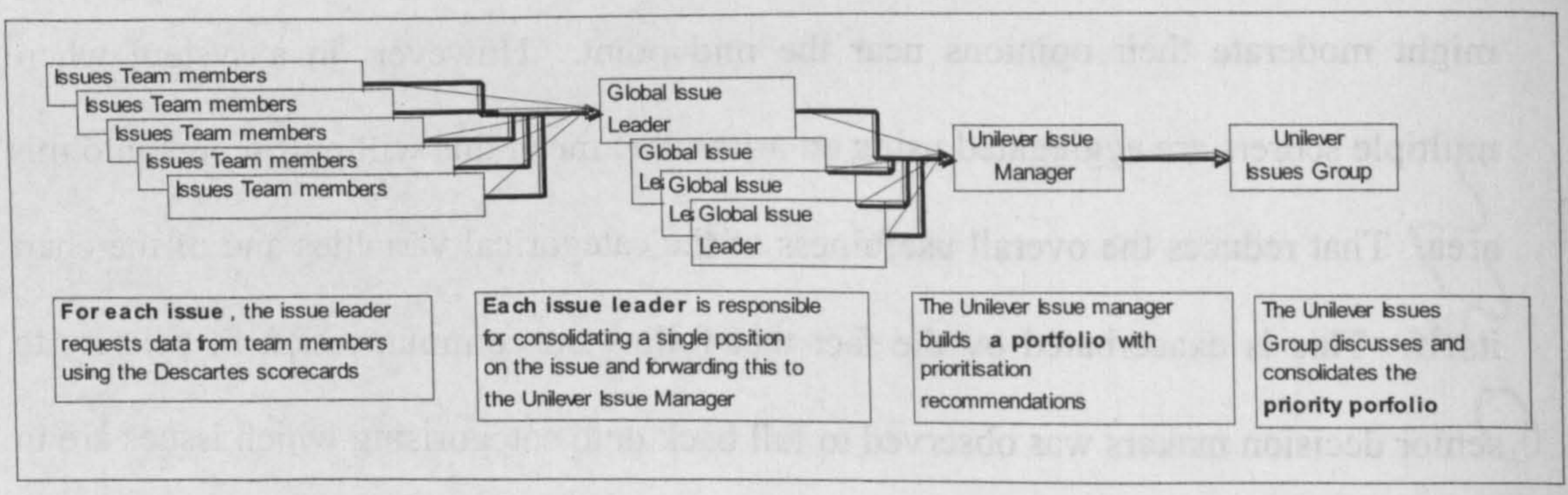
Issue No.12 in the chart below (hygiene and biocides from the HPCE analysis) serves as an example of this. It is at the mid point of both scales and classed as "no real problem". However a half point in any direction would re-classify it. Clearly users might moderate their opinions near the mid point. However, in a system where multiple scorers are aggregated using an arithmetic mean this will pull scores into this area. That reduces the overall usefulness of the categorical variables and of the chart itself. This is exacerbated by the fact that follow on communication of priority to senior decision makers was observed to fall back onto categorising which issues are in which quadrant.

Figure 31 example priority data



Two key developments in the UIG case study would go some way to addressing this problem. The supply chain of data from multiple scorers and the use of time greatly enhanced the usability and usefulness of the graph heuristic. The supply chain of data is shown below:

Figure 42: The issues prioritisation process



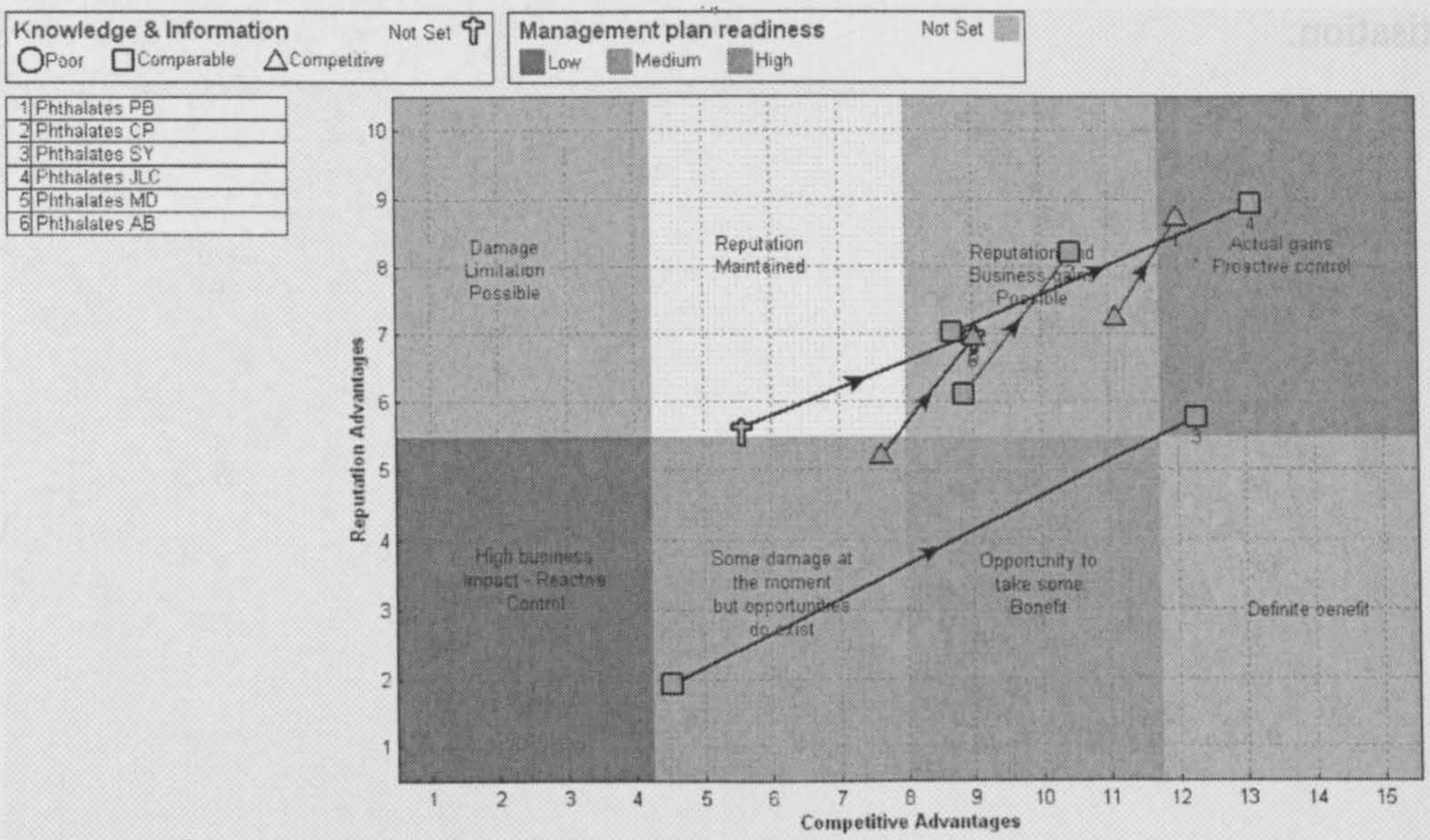
A worked example can show how this supply chain used the graphical environment to preserve reasoning.

Phthalates are a family chemicals used in plastics and in some cosmetics. Chemically very diverse, there are "bad" Phthalates, e.g. proven carcinogenicity, and "good" Phthalates e.g. "plasticers". Put simply, in the global debate, all Phthalates tend to be lumped together leading to calls from Non Governmental Organisations for all of

them to be banned. Very tiny amounts of safe Phthalates are found in Unilever products but the company is heavily criticised. This global issue went through prioritisation.

Phase one: The global issue leader canvasses his team

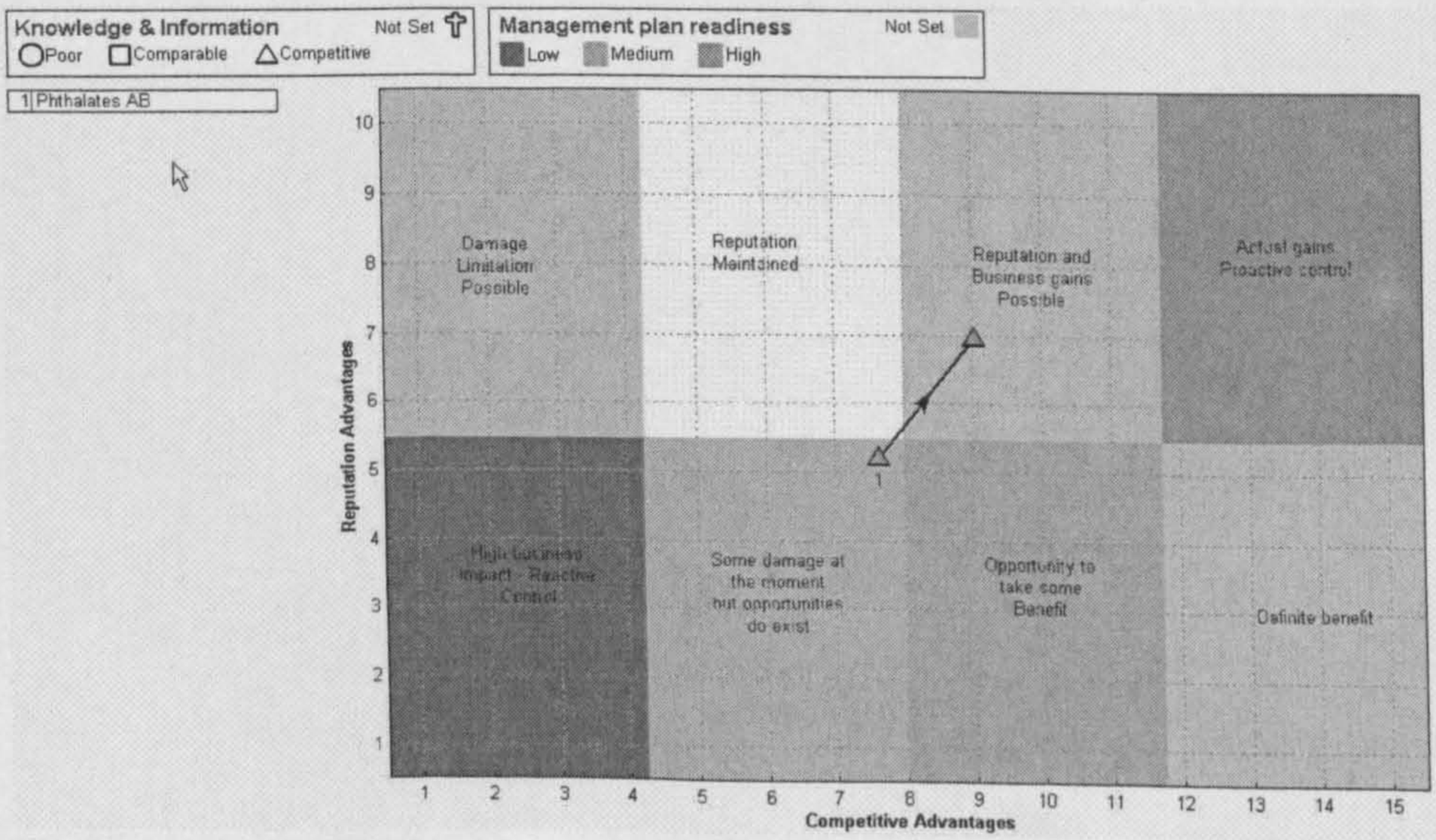
Figure 42: Phthalates



Here, each line on the chart represents the same issue but a different person's view of it. Notice that No.3 (SY) seriously disagrees with the others in the team. This sort of information helps the issue leader reason more fully about the issue in advance of ratifying the issue, i.e. sending the UIG an issue prioritisation report which contains only one set of data.

Phase two: The global issue leader ratifies the team's view

Figure 43: Phthalates ratified



Note here, the issue leader has ratified the data to reflect his own view. This is an entirely legitimate behaviour, but obviously raises an eyebrow. A "history function" of the software allows the previous chart to be viewed from this data. This is the case even where this data is transposed into the strategic score-card. This is a reasoning audit trail which preserves the justification of the priority.

Innovating heuristics

The development of the groups functionality in the ICFE case is very interesting example of innovation brought on by more formal use of these measurement approaches. Here a single data point could now represent many items each made up in turn of multiple attributes. These 'rolled up' variables have three forms of meaning representation, these are

1. When this data point is managed as one aggregate issue on the main plot with 'cross hairs' to connote, but not denote, the variation in both axes.
2. When an independently scored 'proxy' issue on the main plot is related to the content but not any form of aggregation from the underlying scores.
3. When the single issue 'exploded' back into its (measured) constituent variables.

This is a highly flexible form of data reduction operating on a needs basis and only sometimes surrounding a fixed value. That is a heuristic form of reasoning because the underlying reasoning algorithm can vary in detail and form for the same task.

This heuristic is also an important improvement in the prioritisation approach allowing as it does the admission of a great deal more data into the same summary. This, along with the related functionality offered by the 'car park' function, helped the

users adapt to a limitation of their plotting approach when it was swamped with data that they still wanted to use.

4.7.1 Working within, expanding and innovating heuristics

Working within: The early HPCE case had demonstrated that you can work within the 'essential natural heuristics' with which people were creating prioritisation judgements (defining, scoring, analysing and communicating) and improve them. These simple improvements were all accepted and the technology which supported them was "transferred".

These improvements came about not by introducing a model which supplanted what was already there but by introducing a model that mirrored and improved it. The approach retained its heuristic character but these heuristics were now more formal and analytical.

Expansion: What this ICFE case study does however is demonstrate that you can also expand the essential natural heuristics of a user group. You can do this in a reasonably sophisticated manner. Moving the ICFE (and UIG) group of users to a very tightly defined multi-attribute scoring system improved the meaning of their data by far more than the improvements HPCE case. The increased meaning also enriched the quality of the reasoning process which the group was willing to use to apply priority.

Thus the mirroring process in this case helped users to see some systematic drawbacks in the way that they were reasoning. What was clear was that the heuristic could expand to cope. The ICFE community's heuristics for priority, and for handling issues data increased significantly in complexity. However, the final approach was

still proven to be transferable. The community accepted the approach and the technology into their reasoning behaviours. They also reported benefits to that reasoning and its resultant communication.

The result of expanding the heuristics was therefore highly favourable. A group of prioritisation users were successfully taken from a fairly institutionalised Post it pad exercise, one which they struggled to remember the meaning of, to a sophisticated, data driven audit trail of reasoning. Their reasoning became more rigorous and transparent and it could be used to justify the prioritisation of one set of issues over another.

Innovation: In the global issues case the drive for new forms of reasoning (e.g. measuring advantage, looking at strategic impact) meant that new essential natural heuristics could be created out of those presented in the earlier cases. Two of these have already been discussed in the previous section. A third innovation is 'clustering' for decision making.

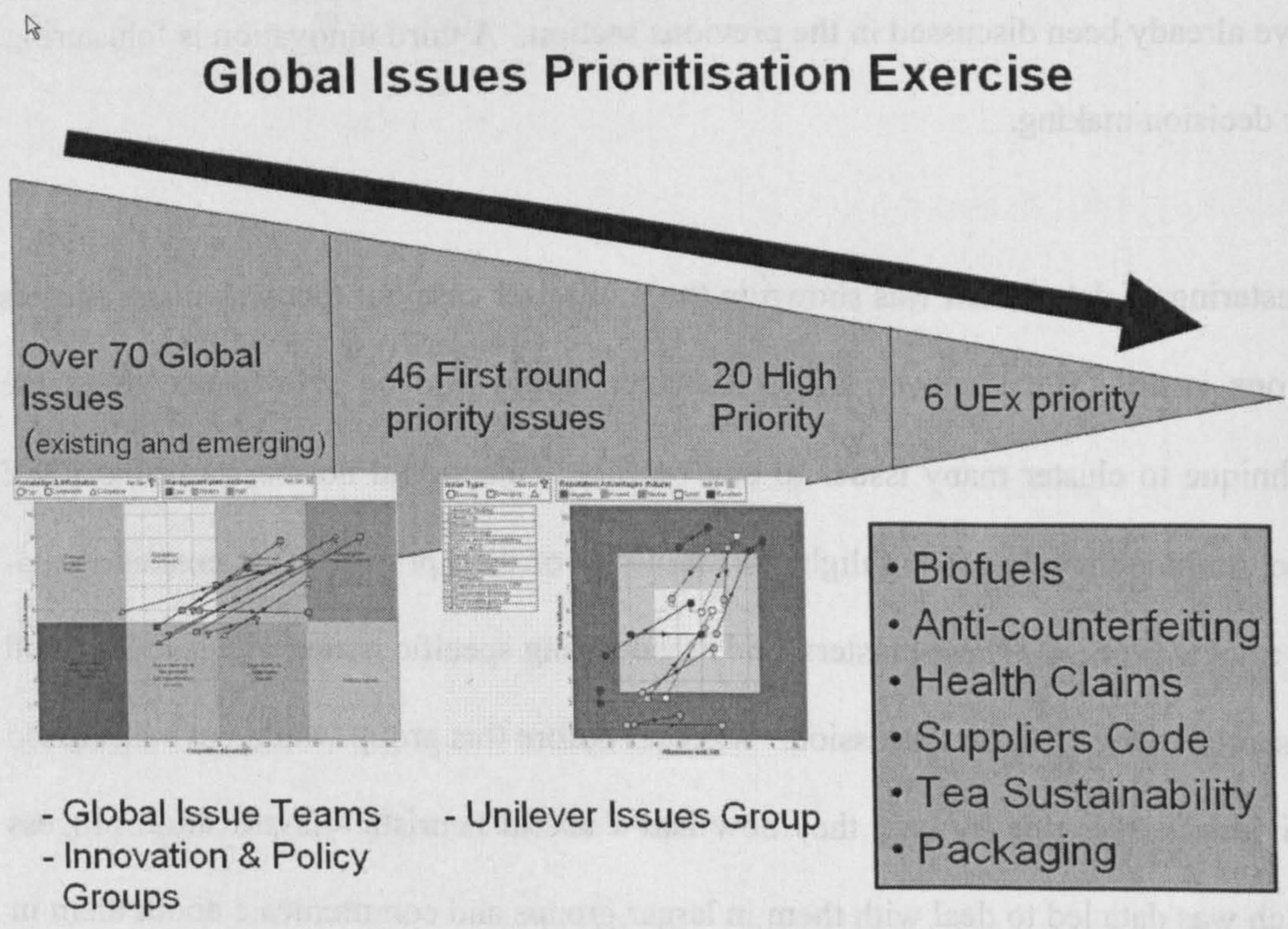
Clustering of data which was shown in the multi-user case but that was many scorers to one issue. The Unilever Issues Manager would also be able to use the same technique to cluster many issues to one profile. This would not be, as in the ICFE case, to hide them but to highlight the similarity of their profile. This creates a data-led review process. These clusters, and any outlying specific issues, are described and presented to the UIG for discussion. Whereas before this group would have discussed individual issues one by one, they now had a useful heuristic classification process which was data led to deal with them in larger groups and communicate about them in that way.

4.7.2 The final global issues prioritisation process in summary

With this work completed, Unilever had, for the very first time in its history, a data driven and detailed prioritisation process. This process delivered a set of priority issues, and the rationale for why they were priorities, based on data. In the final design this was all about "multis" a multi-attribute, multi-scorer, multi-level system. The system as has been shown in this discussion was built upon an evolution of interdependent heuristics. This evolution was based around either essential natural heuristics already in place or innovations of them.

The following graphic was widely used around this time to communicate, in a very simple manner, what had been achieved. In fact, a version of it was presented to Unilever's board.

Figure 50: Global Issues Prioritisation summarised



This case illustrates how all of the essential natural heuristics we found could be improved to generate clearer and more justified reasoning. The scales were improved

to use behaviourally verifiable anchor points. Access to the strategic documents which set Unilever's business priorities created a measurement opportunity about communicating in the language of business. The ability to transpose technical data into the strategic score-card meant this very different strategic rationality was nonetheless linked to the operational world.

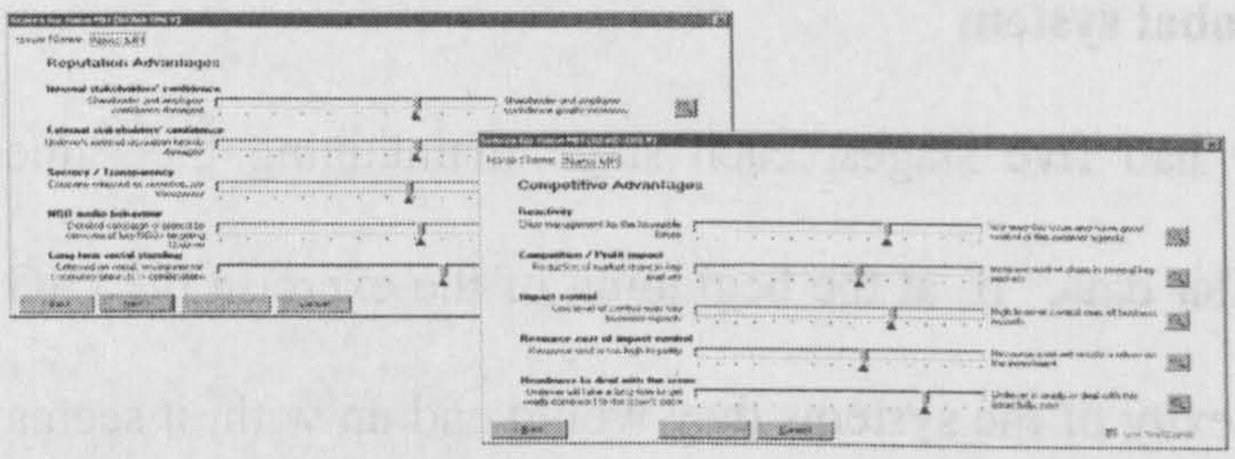
Improved process: the final global system

The prioritisation process now had five stages, each stage maintaining the same rationality and bounded by similar data. If, at the beginning of the exercise the UIG had been informed of the complexity of the systems they would end up with, it seems reasonable to conclude that they would not have accepted something so complex. This is evidence that the journey itself, changed their view of decision support.

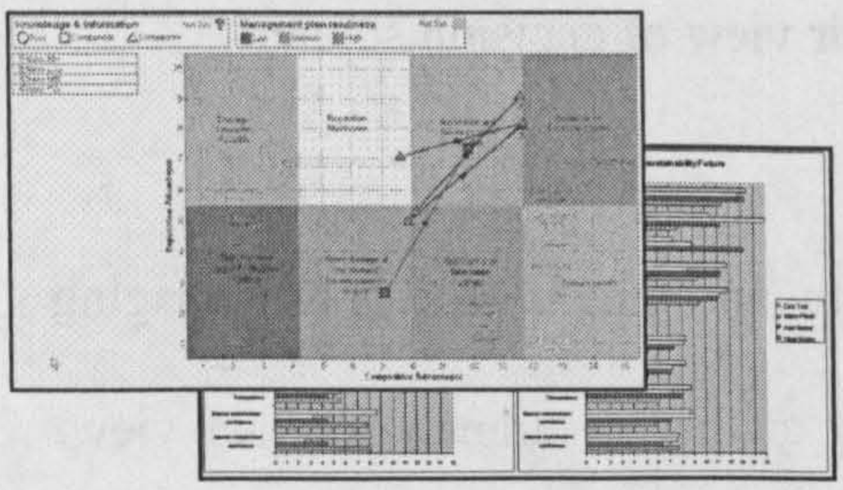
At the beginning of these developments there were two pragmatic worlds of managing issues. The first was from a foods manufacturing quality assurance point of view. The second from the point of view of reputation management specific to the chemicals arena. These were essentially operational case studies. They were testing grounds for the concepts which would be brought to full development over two years.

The final process for global issues prioritisation would generate not one, but two global forms of prioritisation for Unilever. The technical and strategic forms of prioritisation created an unprecedented supply chain of issue data in Unilever, at one time spanning the opinions and scores of over 100 people in one analysis. That supply chain of data was still focussed onto one core concept, discrimination.

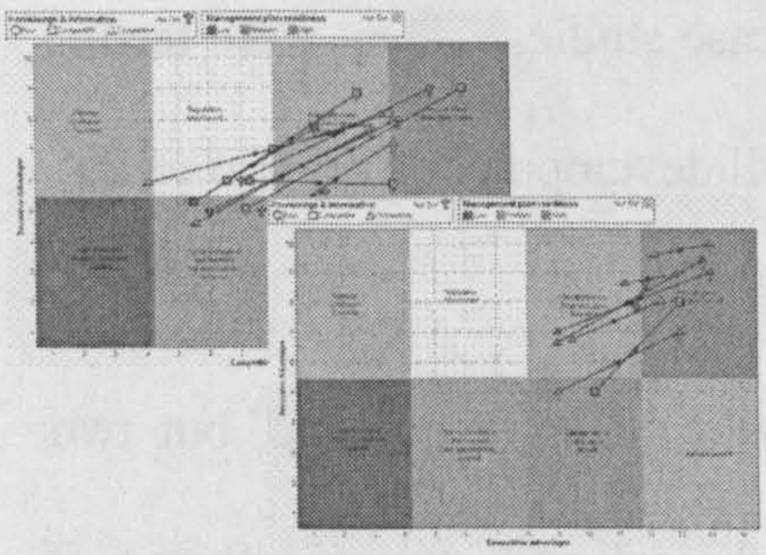
A raft of heuristic conceptualisations representing a range from business criticality to time itself were developed to support this discrimination. These were used successfully to provide decision support to reduce a fiercely complex comparison problem to a single, agreed and accepted proposition. A proposition for use in the issues management teams and in the boardroom. A summary of the final global process is shown:



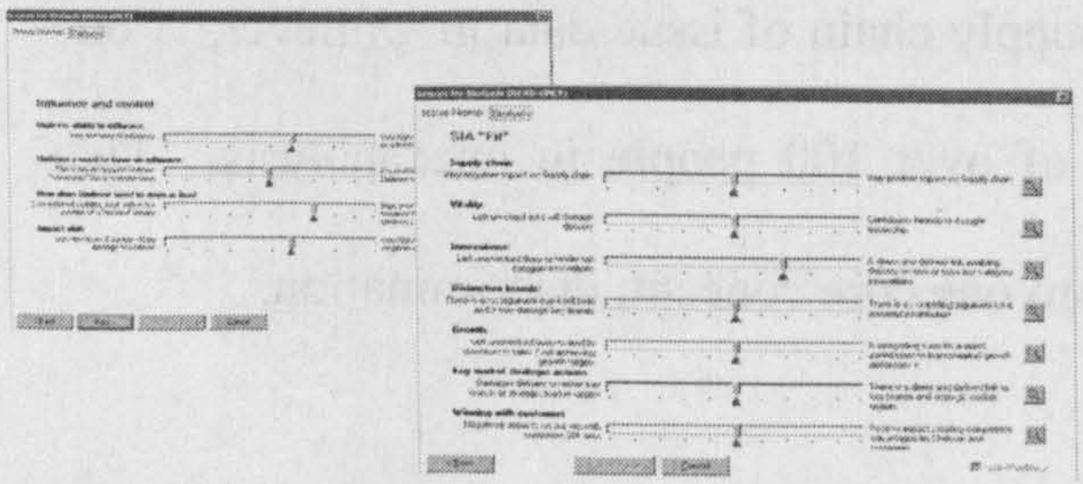
1. Technical score-card used by global issue leaders with their teams to produce and assessment of the current and future state of the **single issues**.



2. Global issue is ratified **to single set of current and future scores** by Global Issue Leader. This is either done mathematically, via the leader's expert judgement or in discussion over key differences of opinion. That discussion may be supported by data from the score-card.

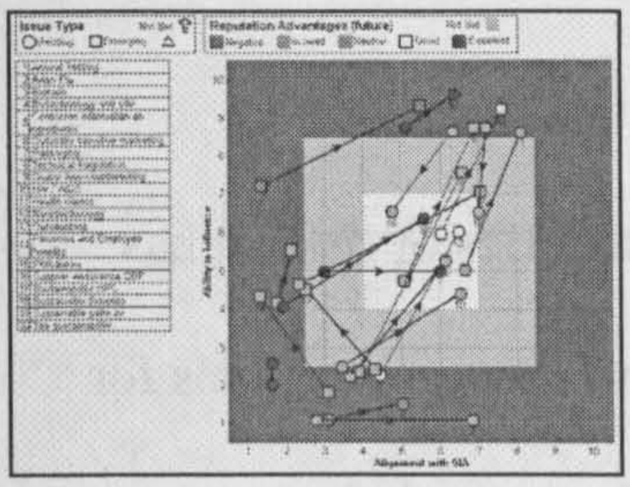


3. The Unilever Issues Manager collates the single issues into **a portfolio** and within that creates **named clusters** of issues with similar profiles for consideration by the UIG. The cluster names contain recommendations for action.



4. The UIG takes a priority list of issues following these recommendations and using the strategic score-card these are now re-assessed for **current and future strategic priority**. Issues which create disagreement are resolved using specific plots

of those issues and, where necessary, a detailed comparison of their scoring profiles among non aligned scorers.



5. This re-assessment results in the strategic portfolio. The Unilever issues manager again assesses the portfolio for clusters of issues with similar profiles. These are ratified by the UIG and recommendations are made to Unilever's board on the highest priority issues and the rationale for priority. The rationale for priority is used as the basis for an action plan to manage the issue.

5. Discussion

5.1 General introduction to this discussion

This thesis has reviewed an applied range of the literature pertaining to human reasoning and decision analysis theories and to the prevalent theories of risk. It has then embarked on a descriptive journey of two heuristic decision support systems for risk reasoning. In that journey I have commented on their influences, their development histories and their application stories. The conclusion of this thesis lies in proposing a marriage between these two manifestly different worlds, that of theoretical risk and decision science on the one hand and that of the "versions" of these one can use in real life on the other.

That marriage faces a number of problems the chief of which is to do with the practical relevance of academic theory to the real worlds of the air traffic control service and the FMCG. The individual case study discussions of these two systems so far has concentrated on their common heritage in a revised, psychologically focussed, idea of heuristic reasoning. That discussion did not contain any consideration of theoretical alignment. This will not be considered in some detail.

The Science of decision

What I believe I have shown is that the 'science of decision' is divided in history and in positioning. It has a lineage which ranges from philosophy through application and back out to philosophy again. The obvious tensions between a closed inventive system of the rarefied mathematics of reason, and the alluring, but "boundary-less" world, of applied reason is plain. There seems to a discernible concern in the literature that one cannot operate in both. The serious tensions between so called normative and descriptive science make this clear. There is also the philosophical

tension over whether heuristics are statistically demonstrable empirical failures or humanist expressions of complex uncharted reasoning forms.

The science of risk

What I have shown is that the science of risk is divided in a plurality of thought forms. These thoughts are trying scientifically, organisationally, institutionally, culturally and socially to say something definitive and helpful about humans and the mysteries of chance, harm and the opportune moment. This discipline has over the last few decades has created a rich, diversified nomenclature and yet has given only a limited a number of validated measurement forms. "What do you mean by risk" is still a chief question and this is after much scientific effort.

I have alluded to the fact that in large, contentious, globally-significant, multi-stakeholder decisions there are highly effective decision models available. These models range from the scientific utility maximising to the social criteria satisficing. We know too that each, from their own particular perspective, is mindful of the value of the other. Maximum expected utilities often contain "acceptability" variables. Social mapping exercises, in the end, contain measured preferences, even if these are unresolved. When there is an awful lot riding on the outcome, governments and institutions call on empirical and social scientists to create equitable, defensible and usable "alternatives" support and, ultimately resolution. These approaches are, of course, very costly. That aside, this is one of the well proven interfaces between science and societal governance.

The science of reason

What I have shown in smaller settings where the decisions are still highly significant (such as NATS and Unilever) is that reasoning is highly valued as the key to safety. Good reasoning preserves an ethical self image story for complex public service institutions like these. This reasoning happens under a complex array of pressures at all times. These are the pressures of modernity. A very large part of that reasoning is technical i.e. is based in the known critical failure rates for complex computer technologies, or in the microbiological science of pathogen death curves in food preparation, or the skin penetration capability of a substance based on molecular size and so on. That part of the reasoning, although not devoid of social and cultural framing, is "a question of science".

When it comes to a risk judgement which is prefixed by "acceptable" or suffixed by "perception" or "communication", and when the reasoning in question centres on values and judgements, the story is very different. In part this is because of an arena shift away from the laboratory, the radar technology and the factory floor. However, there are two main further shifts to do with position of the reasoning in the supply chain of risk:

1. Reasoning is no longer certainly about a fixed objective risk entity, rather it is about larger less definable organisational entities.

2. This concerns, in my language, "upstream" risk, a position predicated on its clarity and on its positioning in the risk decision chain i.e. not clear and very early.

Much of this reasoning, in my observation, is sub-optimally formed. There are two causes for this. The first is that these decisions take place in human group processes, very often around a table with papers on it. The papers often contain a complex

cocktail of present facts and future intentionality. The order of business is often to make a decision or ratify an understanding, but it is also very often to generate, through a consensus process, a model 'of the world of' that decision or understanding. This is a key reasoning form.

The second cause for sub-optimal reasoning is in the facilitated exercises and measurement structures which these organisations have inherited from questionable management science. The flip-chart, the post it pad, the brown paper exercise coupled to the demand characteristics of a decision forum without clear data, all spell uncertainty. The key research question which I feel I have tried to address in this work is: How do you build a decision support system which takes away this sort of uncertainty?

5.1.1 Structuring this discussion

This discussion will be split over three sections, these will be:

1. General discussion of the Hurisk and Descartes systems and their efficacy when benchmarked against the findings of the general literature review.

2. A specific discussion of the possible relationship between the approaches in these systems and subjective Bayes decision analysis.

3. A more detailed look at the systems' main "heuristic devices"

Two key questions have to stay in focus during this discussion :

1. How can busy, and often frustrated, people start to give judgements which behave much more like a deterministic scientific process i.e. they are logical, consistent, transparent and "reasonable"?

2. How can this more logical approach have enough of the characteristics of the existing valued processes that people are already using, to transfer into, and ultimately supplant, these?

The first is a question of logic, it is really for a statistician helped by a psychologist. The second is a question of belief, it is for a psychologist helped by a statistician. Statistics and psychology are the two obvious disciplines to develop a hybrid for decision support in this setting. This will not come about, in my view, by statisticians learning psychology or by psychologists learning statistics. It will come about in a concordat which looks not for commonality but differentiation of efficacy. In a very small way the two projects discussed here are really examples of that sort of thinking.

5.2 Discussion one: of Hurisk and Descartes

5.2.1 Summary of this section

The development of real time measurement and visualisation systems for safety critical and business critical reasoning support is not only challenging, but theoretically is in a very wide world. This research focussed on formalising reasoning processes in two large applied case studies.

These studies had four pillars in common

1. They were detailed attempts to improve the reasoning of groups of decision makers who were depending heavily on traditional methods.
2. They drew on the same core concepts for measurement and visualisation of risk.
3. Purpose built, and deliberately diversified, risk nomenclature was used to guide reasoning
4. Preventing technology transfer failure was the key success factor.

Both studies were founded about two core principles these were

1. Reasoning and rationality should be phrased within the natural reference grammar of an organisation's experts.
2. A reasoning platform for risk, not decisions or expected utilities associated with alternatives was the core object of interest.

The research approach was to look for a kind of supported reasoning where the lessons, factors, insights and breakthroughs for decisions were demonstrated in a more transparent and self-evident system. These features describe decision support system which is:

A reasoning platform approach: Hurisk and Descartes are both reasoning platforms. Hurisk is specifically for risk and strain, Descartes for a risk derivative called issues.

A practical measurement environment: The way in which measurement can be carried out in these systems benefits from the inherent flexibility and usability of these systems where the scoring could be alone, in groups, virtual presence and off-line scoring. The review process could be tailored to specific tasks at different times.

Based on validated core concepts: Concepts derived from the literature to assess the validity of decision support and risk measurement were used to benchmark the systems.

Decision support concepts were : application / workload (the degree to which the reasoning is applied and lower workload); Human centred (the degree to which the reasoning was 'real world'); Rationality (the degree to which the approach and algorithms would satisfy questions of scientific rigour) and Values / credibility (the degree to which the systems were valued by users as credible).

The evidence in the study suggests that the systems offer good decision support because users can be seen be using and valuing them and acting rationally on the basis of their output.

Risk concepts were: application (the degree to which risk is about measurement communication, decision making etc.) and perception (the degree to which can convey multiple perceptions and cultural importance of risk factors etc.)

Hurisk is shown to be a comprehensive decision support methodology with a coherent design which requires thorough execution. The system functions as a decision transparency aid supporting risk based decision making. The value of Hurisk to NATS was plain in identifying and scoring non technical risk as an urgent priority

was demonstrated. The definitions of risk allowed for an affect component which is enshrined in the organisation's commitment to safety. Further research is needed to look at the long term effects of the use of the tool.

The conclusions of this section are that the design of these two systems is fundamentally sympathetic to the organisational behaviours of their hosts. This has created a powerful level of technology transfer into these environments. Where these systems are inherently strong is in their ease of use and the familiarity of their heuristic architecture with that of the users. Where they are potentially weak is in how transportable their heuristic credentials are in terms of measurement consistency and efficacy, on the questions of bias reduction and/or introduction. Finally whether more statistically oriented elements might strengthen them whilst having them remain a transferred and valued suite of tools is a key research question.

5.2.2 Some key points raised in this discussion

The key points raised in this discussion fall into one of five categories, these are:

1. General assumptions from the theory base
2. General observations about the benefit of heuristic methods in these designs
3. Benchmarking heuristic systems against standards that can be derived from the literature
4. Similarities and differences between a heuristic approach and a subjective Bayes one
5. Technology transfer issues

General assumptions from the theory base

- An applied form of decision support for the communities studied in this thesis lies at the intersect between decision support and risk.
- The science of decision support is complex and unresolved resulting in a number of overlapping candidate designs for support tools
- The science of risk is predicted to remain lacking in validated measurements due to its disputed nature.
- The use of the term heuristic may be philosophically confused due to an over reliance on one group of decision theorists' use of the term.

General observations about the benefit of heuristic methods in these designs

- The nature of the heuristics in these models is explicable by looking at the metaphors which they evoke and understanding how the heuristics fail to satisfy the necessary objective qualities of the target object.
- It is possible to model decision support inputs further upstream than is conventionally done and reflect nature of the uncertainties as more granular.

- It is an important task to try and remove the uncertainty associated with the sub optimal decision processes observed to be operating these settings.

Benchmarking heuristic systems against standards that can be derived from the literature

- In heuristic reasoning about risk users can be shown to place a high value on the narrative constructs this uses. This concurs with recent theory on heuristics and biases which suggests that narrative (or semantic) approaches reduce bias.
- An assessment of a good decision support system based on a summary of the extant research produces eighteen attributes which can be divided into four key categories.
- Comparing the systems in this project with this list shows that the systems are conceptually very well underpinned.
- The same exercise for risk theory delivers a less articulate idea comprising of nine attributes across two criteria.
- Assessing the systems against these demonstrates that they are performing modestly, in part this is to do with the vague way in which risk can be appropriated like this.

Similarities and differences with subjective Bayes

- In large, highly significant, multi-stakeholder decisions the use of very complex models, such as those produced by a full subjective Bayes approach, will be considered as cost effective in terms of time and intellectual overhead.
- In small scale, high volume decision problems concerning risk something less analytically and computationally intense will always be required.

- Heuristic reasoning systems arguably provide a more direct form of access to the reasoning objects in use, this is much like a (simplified) version of a Bayes model output without the overhead to produce it.
- The algorithms in these systems seem to form around an idea which is like a simplification of subjective Bayesian decision modelling designed from a primarily psychological standpoint.
- A hybrid systems between heuristic reasoning and subjective Bays is suggested to deal with rejection, by non mathematical users, of Bayes on the grounds of complexity.
- The key similarity between subjective Bayes and heuristic reasoning approaches is a move away from providing an accurate answer in favour of an accurate description of what is important and why
- A key difference is that in a heuristic reasoning system it is considered enough to have laid bare the rationality with higher level upstream measures which remain available for later attention.
- Heuristic reasoning systems arguably identify important but generic aspects of the problem spaces. These do not have to form a part of a unique model as they would in a Bayes style approach.

Technology transfer issues

- Conclusions drawn from this research demonstrate that heuristic reasoning concepts have high validity with users, are inherently transferable between decision problems, and are generic, quick and effective in a way that other decision models cannot be.
- The result of these strengths is that these systems have penetrated the organisations to a significant level where more complex modelling forms have not shown a track record of being able to do so.

5.2.3 Discussion

As is evident this thesis lies at the interface of applied occupational psychology, risk analysis, decision support and human reasoning. The development of real time measurement and visualisation systems for safety critical and business critical reasoning support is not only challenging, but theoretically is in a very wide world. In applied terms I have narrowed this considerably. This is both to take into account what was feasible to achieve for real world sponsors, and to avoid a post-modern vortex.

This work is about formalising reasoning processes. These detailed, often very soft, but definitely very applied, case studies looked at two different organisations, aviation safety and consumer goods. The case studies also looked at two very different (on the face of it) reasoning objects, risk-and-strain and issues. What these studies had in common can be described in four pillars.

Pillar 1: Both were detailed attempts to improve the reasoning of distributed groups of decision makers who were depending heavily on traditional, consensus-based, boardroom methods.

Pillar 2: The Unilever case study was a natural extension, and then evolution, of the proven reasoning forms from the NATS work.

Pillar 3: Groups in both organisations, once they were using the methods we designed, were reasoning about risk under a newer and deliberately diversified nomenclature. (An issue, a strain, a decision influence, a lesson, in all cases these would have, in a more rudimentary approach, just been called risks).

In these frequently very tough-minded organisations the challenge of providing better decision support had to be proven through organisationally effective measures. This is why both case studies tell a story of the indigenous process to develop decision support from within the ideas the organisations were already having on risk monitoring and tracking. Even if this resulted in the rejection of some of these ideas, this is the first foundation of this research approach:

Reasoning and rationality should be phrased within the natural reference grammar of an organisation's experts.

Risk in these studies is an "upstream" concept. Risk in Hurisk may be well defined in a multi-attribute scaling space. I may be well presented in a complex array of utilitarian comparisons. However, it is still being intercepted at "pre-decision" point. This gives a second foundation of this research approach:

What is needed is a reasoning platform for risk, not decisions or expected utilities associated with alternatives for risk.

The audit trail for decisions is a crucial example of this approach. Both systems described here have detailed history features where experts' reasoning can be laid bare in some detail. It is in this supported reasoning itself where the valuable lessons, correction factors, insights and breakthroughs for decisions were demonstrated. This is decision support, not decision analysis.

Heuristics and their application is the third foundation in this research. In the first instance these had to be recovered from a very simplistic, or single minded (but relevant) application. Heuristic reasoning in the lingua franca of decision theory had

become synonymous with systematic errors and irrationality. This thesis has contributed to a reclamation of the term as a working concept. This work has proven convincingly that deliberately applied heuristic reasoning methods have a great deal to contribute to business and safety critical industrial problem solving.

Pillar 4: The fourth and final foundation of this work is the prevention of "technology transfer failure". This is a technical term from Human Factors often applied to device penetration into an organisation. Technology transfer relates to the factors which prevent inherently successful devices from being taken up. I've broadened that idea by using "technology" to refer to decision processes and tools and describe how these have influenced the behaviour and the culture of the host organisations.

Because of the upstream nature of the risks and processes in this research, technology transfer success was approached through building a system which was 'for experts', not by experts. A system which was not, in and of itself expert. 'For experts systems' rather than 'expert systems' might seem a subtle distinction, but it is key to understanding why this research took the direction it did and it remains my main conclusion that this approach is highly effective.

Developing 'for experts' systems

My starting assumption was that a close affiliation with practitioners and their working environments would be able to understand what sorts of risk and decision support models would be needed. That is why there is a heavy emphasis on social research techniques in my methodology. I wanted to reflect back to these organisations decision support systems which were in what I have called "their own reference grammars" for risk and decisions. The resultant systems faced in a

theoretical direction where I believed mainstream decision support was not being used effectively. The application of decision analysis and support to organisations, as I had observed it, was too focussed on the idea of external experts building complex and expert systems.

A reasoning platform approach: Hurisk and Descartes are both reasoning platforms. Hurisk is specifically about risk and strain. Risk is assessed through the development of risk events and the association of three scores to these. A probability that the event will come to pass. An impact on a project or an intention, should the event come to pass. An estimation of how controllable either the probability of occurrence or the mitigation of impact would be.

Strain is assessed by use of one of two possible questionnaires (a longer and a shorter version). The questions relate to real states of affairs. Users agree on the degree to which, whether in the particular project or status of an operation at a given time, these states of affairs accrue. Collections of these states of affairs are in strain categories. These categories are have their relative importance estimated using a weighting structure. Visualisations support the review of this data over time, and between users.

Descartes is about a reasoning platform for issues. It has taken the essence of Hurisk, the tiered scoring and visualisation of risks, and simplified it in one sense and made it more involved in another. The simplification comes about in the level of reasoning. Issues are larger collective ideas than risk events. Ideas which can have no meaningful probabilities associated with them which is why Descartes does not use this measure.

Complexity in a Descartes style reasoning task comes about in three ways. First, issues are assessed for their status both now and in the future. The discrepancy is visualised as a delta value which is an input to the reasoning task. Second, issues are assessed across a large group of users, 111 was the largest. The reconciliation of the different stories these assessments are telling is part of the reasoning task. Third, issues are scored across two core propositions, a technical de facto importance and a strategic relevance. The quantitative data from the former is a selection criteria for the latter.

In NATS and in Unilever two reasoning tasks are being approached by this system. Groups of users apply these two systems to decision making tasks e.g. what are the key project risks, what sort of strain is the organisation capable of absorbing before safety is eroded, what should be done about high priority issues and what does that priority mean in terms of business opportunity or threat now or in the future.

A practical measurement environment: The way that the systems are applied follows a standard research metaphor. A sample of key people are identified, a set of independent (or group) scores is generated. The sample scores are compiled and comparisons are made to generate insights. The analysis is repeated a number of times and results contrasted over time and between groups / individuals to generate further insight.

Hurisk and Descartes were purposely designed to be heuristic systems. Two key ways in which this was achieved will serve to illustrate this at this point. The two examples are definition of the core entities and the behaviour of the measurement systems.

User-centred definition: The definition of a risk in Hurisk was controlled by a set of seven rules which users had the choice to satisfy or not. Strains were pre-defined through research but were very fluid at the interpretation stage. Issues definitions in Descartes were user selected (and very wide ranging). The items entered into these reasoning platforms were more typified by dissimilarity. This dissimilarity was irrelevant, what was required was a currency to compare them and common visualisations to consider the comparisons dynamically.

Encoded measurement: The measurements used in Hurisk were fixed to more definable entities. However, the algorithms which governed the final scores were psychological in character. The 'knowledge score' associated with probability, control and effect was an uncertainty co-efficient. It formed a psychological narrative about personal belief calibration but this was not an actual calibration. Likewise the risk tolerance algorithm mediated the extent of uncertainty but this was controlled by a variable user setting over time. That setting was a psychological narrative to do with feelings about accepting risks or not.

In both Hurisk and Descartes all measurements could be performed at multiple levels i.e. ranging from physically changing a position one data point to providing views on over a hundred questions. Whatever level of scoring the user chose to use, the system made sense of their view by trying to compensate to maintain any detailed profiles which previously existed.

An idealised rationality: The benefits from the inherent flexibility and usability of these systems was to reasoning. Users were able to ask these questions over the full

gamut of reasoning environments they would normally use anyway (alone, in groups, virtual presence, off-line scoring and review). The benefit of both systems was clearly not about making a calculation, this was secondary. The benefit was released because each system had been designed to embody in a flexible shell, the idealised rationality of the decision makers.

This idealised rationality was released to a greater or lesser degree depending on scoring diligence, time and other operational factors. Importantly all of these factors were outside of the systems parameters. The system focussed on asking questions and making comparisons (over time and change). The ease with which the scoring activity was sublimated to being representational, not deterministic, helped the users to use their measurement results to tell stories. This led, in the users' views, to better more informed decisions and better risk management.

Summary characteristics

In summary both systems were soft reasoning platforms with hidden hard algorithms. They could reason about risks or issues. These risks and issues were measured heuristically as a reflection of belief and judgement allowing for bounds to be stipulated and possible changes over time to be reflected. The scoring systems ranged from the crude to the highly sophisticated within a single housing and were applied at the users discretion with underlying cues as to which method had been used. These cues, algorithmically, created a system preference to profiling more complex data forms even in the face of crude over-writes. The scoring process was distributed across people, locations and times. The measurement criteria were forced whether the objects were commensurable or, as was more often the case, were not. Uncertainty

was coded either as bounds, over time or between judges and at all times held in tension.

Three things stand out in these systems: core logic; transparency; and representation. The core logic was that the objects were interrogated by organisationally meaningful questions leading to enhanced understanding and dialogue. The system provided transparency at every level of scoring (and resolution) to create audit and justification of reasoning. At all times the systems provided a single reduction of the data into a visualisation of its status, thus coupling a heuristic realisation to the idealised reasoning.

5.3 Do these heuristic systems offer good decision support?

The argument from the evidence so far is that my systems offer good decision support because users can be seen not only to be using and valuing them but acting on the basis of their output, often without further introspection. That has a high level of face validity but it doesn't necessarily address any idea in the theory.

As I maintained earlier the problem here is really one of definition and, given the range of often conflicting definitions for decision support and analysis and their theoretical emphases, the world of decision analysis itself might not easily answer that. However, having reviewed the various schools and methodologies we were able to discern an idealised qualitative requirement set for decision support, it looked like this:

Core concept	Eighteen features of an ideal decision support system
Application / workload	Can address small scale few stakeholder problems
	Does not have too many variables
	Is not highly demanding of time and resources
human centred	Phrases propositions in favour of long term memory use
	Rests on semantic and narrative descriptions of problems
	Takes cognitive limitations into account
Rationality	Mirrors accurately what people really do but in rationality terms
	Is sufficiently comprehensive
	Uses weighted subjective criteria and objective
	Includes the best available mathematics suited to the problem space
	Involves some formal logical rules for decision closure (choice, acceptance)
	Bases decisions on a bridge between induction and deduction
	Is able to identify and counter biases in reasoning, or at least inconsistencies
Values / credibility	Has credibility with users
	Allows decision making to take place under conventional circumstances
	Remains transparent to and explicable by decision makers at all times
	Allows individuals to compare reasoning on a common platform with each other
	Reflects explicitly the values that are being brought to bear

In the following section, using one example per item taken from both systems I will argue that these systems show many attributes of being sensitively designed to offer high quality decision support.

5.3.1 Application and workload

Can address small scale few stakeholder problems: Hurisk and Descartes can work with one person, one set of data and in cases where this has happened the users have reported benefits. They can of course also work with user bases in triple figures but the design is essentially for the small group resolution process both in terms of risk and issues profiling but also in management decision making.

Does not have too many variables: Descartes has two variables, so does Hurisk. The fact that these are backed up by multi-variate spaces and multi-attribute scaling does not weaken this ideal. This is a lean measurement style, in both cases to perform a robust assessment requires only two numbers.

Is not highly demanding of time and resources: The maximum times for analyses of fixed numbers of issues was almost written into the computer code of these systems. They were measured in minutes. However, the systems were designed with a siren like quality to lure the user into performing more detailed analyses which take very much longer. Being highly demanding of time and resources is of course a cost benefit argument. In comparison with complex decision analysis systems Hurisk and Descartes would still come out well.

5.3.2 Human Centred

Phrases propositions in favour of long term memory use: This is a very interesting criteria and one where Hurisk and Descartes show different and possibly complimentary strengths. In the Hurisk strain questionnaires there is a case for suggesting that these are a best practice when it comes to "corporate memory". The invocation of strains from the past are classified and grouped in a way which is indicative of long term memory and evokes a long term memory response in the analysis. Likewise risk in Hurisk is a live form of recalling from the past what can and did go wrong and applying it to a future proposition, this too is reasoning in that part of the cortex which draws on remembered experiences and reactions to those of others. In Descartes the objects of reasoning are quite simply long term in their nature. They can only be understood in terms of stories from the past and projections into the future. People in both these systems are working on long term artefacts.

Rests on semantic and narrative descriptions of problems: Risks in Hurisk as phrased as events, strains are phrased as potentialities and possibilities. Each is authored either by people from the NATS culture or in the language of that culture at the time of the original research. These entities are language forms they are not technical or mathematic propositions. The problem of risk and strain however, is not formally described. It is clearly present as a loose narrative but it could be improved. In Descartes there is again a core narrative, if one 'reads down' the side of a score-card one can see a worst case scenario for a problem keenly exemplified. This too however is not quite what the statistical theorists are saying here. Partly neither system is doing this because they are aimed at upstream risk and the downstream decision is still left in the hands of subject matter experts.

Takes cognitive limitations into account: None of the heuristics and biases ideas or the wider cognitive psychology in this case has been formalised into these systems. This is a key place to criticise them. It is important to comment on data from users about how helpful these systems are to their thinking, but there is not yet any formal evidence to compare this with alternative methods or look at things such as measurement consistency etc. As users these people are in applied communities and it seems unlikely that this sort of approach will concern them.

5.3.3 Rationality

Mirrors what people really do in rationality terms: In both systems the use of the four quadrant plot as a simplified visualisation is an exact copy of what was happening in their boardrooms. My considerable improvements to its functionality (expanded aspect ratio to reflect positive and negative zones side by side, the use of a boundary argument for strategic issues, the introduction of between user distance plotting for risks) all sought to take advantage of the benefits which people were getting, or perceiving, from this approach. The addition of real questionnaires, formal in depth scoring and multiple comparison options greatly improved the power of users to interrogate the kind of data which they displayed in this way in a live setting.

Is sufficiently comprehensive: Hurisk's comprehensiveness is obvious from two examples. First, benchmarking in NATS showed it to be unique as a thorough multi-attribute treatment of risk with as thorough a measurement scale attached. Second, the strain questionnaire was the result of exhaustive research and testing with key users and was a landmark piece of work just on the number of items in the checklist alone (in some settings this would not be good but in aviation safety the checklist is very highly vaulued). The fact that it reflected a balance of well known technical risk

sources but, for the first time in that culture, actually introduced psychological and emotive reference points makes it clearly a comprehensive piece of work.

From the Unilever perspective Descartes was comprehensive because for the first time ever in the company the entire issues portfolio (over 100 issues) was analysed by the entire issues community on one platform and one commonly agreed score-card. The tool itself could then be used to communicate throughout the organisation the results. There were very few examples of this kind of tool available even to one of the worlds largest companies.

Uses weighted subjective criteria and objective: In both cases the obvious focus of attention lay with weighted subjective criteria, this was the heart of these systems being heuristic. However, that does not underplay the presence of objective data in the reasoning frame.

Unilever was keenly interested in cost and cost of control and, where appropriate, actual costs were part of the analysis. The position statements which were used to frame the issues definitions which were used within Descartes were, in the technical cases, the results of the best available science. In the more subjective cases they were still strictly argued positions of fact. For NATS the purpose of this particular system was not to be objective but to be complimentary to existing objective systems.

Includes best available mathematics suited to the problem space: Both of these systems sit upon a very basic combination algorithm which propagates a multi-attribute score forward to a single score and vice versa. Attached to these are the algorithms described above. It is uncertain if one can comment on these algorithms

being the best available. They were designed from the psychology upwards. As one of the conclusions of this thesis is that they could and should be more powerful, we have to conclude that the maths in question, although proven to be suited to the problem space, is a work in progress.

Involves some formal logical rules for decision closure (choice, acceptance): This is not the case for these two systems. The rules for decision closure remain the same as they were before these systems were invented. Their aim is to support.

Bases decisions on a bridge between induction and deduction: Here again we have an uncertainty. What is really going on when Hurisk and Descartes users make their judgements. One could argue it is a kind of glorified coding of hunches. Or is it a fine tuned, stylised measurement system tuned into to highly intangible variable sets. The thrust of being a fully heuristic system is that they exemplified in the fact that the data quality is entirely in the hands of the users, not determined by the system. Only the users can hope to judge their value therefore. Their accuracy is also a moot question as the logic of the reasoning that people are using is not consistent and therefore hard to define standards for. Insight is of course a product of both induction and deduction, but I have not identified when either is occurring.

Is able to identify and counter biases in reasoning, or at least inconsistencies: Hurisk and Descartes, as has already been discussed under other sections above, were designed to do this as a primary output and each has a complex array of ways to track and highlight reasoning and reasoning agents' profiles over time.

5.3.4 Values / credibility

Has credibility with users: These two systems both achieved credibility for opposing reasons. In the NATS case a detailed user testing process noted that the in the reasoning objectives of the system the thing that was highly valued was "it really makes you think". In the engineering case studies the system replaced a standardised risk assessment which again suggests that it proved its credibility.

In the Unilever case, this kind of ad hoc reasoning culture the system had most credibility because it was quick, easy to use and a discussion stopper. The thing that was rated very highly was the ability to "get to the point". This might not satisfy a measure of adequate reasoning but in in terms of adequate risk reasoning I might question that as a credibility measure, but it was certainly highly prominent. I think however the obvious credibility that Descartes achieved was that it was used and is still used by the executive and the board of Unilever to guide their priority setting. In this sort of tough minded world with often very low tolerance for ambiguity this is a real achievement.

Allows decision making to take place under conventional circumstances: In both Hurisk and Descartes the decision making still took place in exactly the location it had been in before with exactly the same groups of people in control of the decisions. The difference before and after was that in one case they were discussing and using consensus based techniques and in the after case they were visualising, debating, profiling and in some cases rationalising real data in a live system. The outcomes, very much the focus of both systems were still left outside of any system output. Both Hurisk and Descartes have extremely tightly bounded definitions as reasoning

support. They can only be used to help decision making, they do not provide decisions.

Remains transparent to and explicable by decision makers at all times: As has already been described the power of both of these systems is in their use of multiple highly flexible environments to create the space for disputation and focussed thinking. This is the key platform for their positioning as group reasoning tools. In Hurisk, you can track your risks and strains from the moment you started recording them, you can analyse your own and others scoring styles and preferences, you can compare trends in your own and others data. In Descartes the history of the issue's scoring (a key factor in the decision transparency argument for such complex objects) is available throughout the analysis even when score-cards are interchanged. It is not only possible to see why an issue passed the first stage of priority it is also possible to see all the individual scorers from the first generation of scoring. Hurisk and Descartes are by design transparency enhancing systems because lack of transparency in the Unilever case was felt to lead to bad judgement and in the NATS case it was considered dangerous.

Allows individuals to compare reasoning on a common platform with each other: The argument above also addresses this point

Reflects explicitly the values that are being brought to bear: This is not so easy to lay claim to. Certainly the use of measurement metaphors (clock faced dials for NATS, sliding colour bars for Unilever) reflected what kinds of measurement had value to the audiences. The inclusion in the Unilever work of issues scored as both risk and opportunity reflected the value tension of the two cultures who wanted to use

the system in different ways. The idea of priority represents some kind of value metric.

However, these would not seem to compare well with, for example, Keeney's idea of "value". Within the Unilever work however, if you look more deeply at the scale items in the reputation and influence scales these are clearly statements which expose the company's self perceptions as a good ethical corporate citizen. In that way values are expressed. The chief value being expressed in NATS however is at the meta level of safety being important enough to develop and use Hurisk in the first place.

5.3.5 How did the systems compare?

Hurisk and Descartes, if the list of eighteen attributes were actually a scale, would score in the region of 12 out of 18. Many of these eighteen variables (and decision analysis approaches they purport to) seem to be suggesting it is important to control key aspects of the decision reasoning space which Descartes and Hurisk do quite well even though they are only heuristic systems. The fact that heuristic systems help with decision framing, reasoning and so on, rather than decide puts them at an advantage.

For experts systems match up rather formidably with a Gestalt of good decision support. That, of course does not take away from the fact that formal decision analysis systems have to come up with an answer and decision support systems do not.

5.4 Do these heuristic risk measurement systems offer meaningful risk?

In a similar way one could ask the same question about our qualitative attributes for risk.

Core concept	Nine features of an ideal risk system
Application	Risk is about measurement and management
	Risk is about communication
	Risk should be able to inform actions and influences
	Risk is about decision making
	Risk has to be about control features
Perception	Risk has to convey multiple perceptions
	Risk ownership and risk framing are key essentials
	Risk should focus on culturally important things
	Risk should contain a value idea and have an affect component

In the interests of economy I will confine my comments to the sphere of Hurisk, which is the system formally aimed at risk (including the idea of strain). There is an argument for issues is Descartes being a larger cousin of risk. The answers in terms of the quality of the risks used within the system generated is conjectural. What we will do is compare an ideal proposition, (the seven risk rules, and a fixed risk measurement scale) with the attributes above.

5.4.1 Application of risk

Risk is about measurement and management: Here Hurisk clearly scores high being a dedicated risk identification and measurement system. In Hurisk, the definition of risk is very clear and the conceptualisation of risk is multi-attribute. The use of Hurisk leans heavily in the direction of 'guided thought' for risk remediation i.e. understanding why a risk is a problem, be it control or impact, be it the levels of knowledge available etc. Strains too are measured and in a more detailed manner, given a pre-existent and highly valid cultural model for sources of strain under-

pinning the measurement. Any problem or disagreement can be readily and quickly highlighted. Taken across the attributes already described Hurisk is a very effective measurement and management tool.

Risk is about communication: The purpose of Hurisk is transparent reasoning about risk across users and over time. It's outputs are all designed to build better risk narratives graphically and in the changing differentials of people's beliefs about risks and strains. These are all designed to be taken back into the group processes and shared as a controlling story for future actions.

Risk should be able to inform actions and influences: Again here, because the data quality is in the hands of the user, it is hard to say. The efficacy being introduced is again in operational responses to the presence of the tool and to the data it gives. In terms of detailed risk and strain profiles Hurisk certainly provides a handsome narrative, but it does not control how that is used in related decisions.

Risk is about decision making: Hurisk is a decision support methodology it has a comprehensive design and requires thorough execution. However, it remains a support nonetheless. It is assumed that Hurisk's influence on reasoning day on day is good. It does produce a deliberate transparency to create risk sensitised disputation mechanisms. The risk and strain lists it creates are all going to aid good decision making. Having these assumptions even with endorsement from the users is not the same as proving them though.

Risk has to be about control features: Hurisk prioritises risks over time and between and within users as well as running an overall risk burden measure. It allows

users to stipulate their perceived levels of risk tolerance, a vital pre-cursor to any nuanced control strategy. Moreover, Hurisk, of course, requires users to make an assessment on a risk by risk basis of their control.

5.4.2. Perception of risk

Risk has to convey multiple perceptions: Many of the comments above hold sway here when crossed with the idea that Hurisk is a multi-user tool. All users in a project have access to all data and there are deliberate and sophisticated comparisons over time available. This even extends to the levels of measurements that individuals are using and with what frequency and in what scoring style.

Risk ownership and framing are key essentials: Here Hurisk has no obvious strength, the risk elicitation process, although guided by the seven rules, is not under the control of the system. The social processes involved in the framing of the risks themselves, again helped in accuracy perhaps by the rules and the possibility to "splice and join" risks and their scores once they have been defined, is not explicitly controlled.

Risk should focus on culturally important things: Care needs to be taken here to differentiate between internal and external cultures. Internally Hurisk is highly validated. First and foremost because strain items in NATS come directly from the culture, second because the entire system was designed inside the culture and thirdly because the metaphors and the reference grammar are highly culturally specific. Having said that, one source of this attribute for risk was in reality speaking about different national cultures against which Hurisk has not been benchmarked.

Risk should contain a value idea and have an affect component: The value of Hurisk to NATS was plain from its originating research, there were other sources of risk in the world than technical ones and a system to resolve these was an urgent matter. The affect component is found in the commitment to safety. Hurisk demonstrates this in its explicit requirements for a risk tolerance measure to be added. This is however a rather weak argument compared to what this literature was saying about switching between modalities and their relevance to guide decisions.

5.4.3 Does heuristic risk meet the standard?

Again using the notion of scoring on these nine attributes, Hurisk scores five. This is just over 50% and that is not as good as it could be. The answer lies in further research to look at the long term effects of the use of the tool and the benefit of that data will not be available until there is further research. The answer also lies in the positioning of Hurisk as a heuristic system. To entertain the lassitude this system creates to be used in uniquely creative ways which leave the user determining their own course in the decision means they will do just that. A more formal system would be able to highlight where this is positive and where it is negative and provide further inputs to support. Hurisk and Descartes cannot do that.

5.5 Conclusion to general discussion of Hurisk and Descartes

What these two systems show, when you compare them as best you can with the stretching, and hard to articulate, demands of extant theory on decision science and risk, is that they are strong where they are strong. Their design, fundamentally sympathetic to the organisational behaviours of their hosts, has created a powerful level of technology transfer into these environments. In the case of Hurisk, Unilever has even started to tentatively use it now for project risk management and that is an interesting cross-over (because they previously rejected it as too complex). In the case of Descartes, due to an insightful design decision early on to make it a generic, it has, at the time of writing, five distinct applications in Unilever. It is currently being rolled out across the globe as best practice.

Where these systems are weak is of course now the interesting focus. Are their heuristic credentials in terms of measurement consistency and efficacy a compromise in terms of understanding their impact on real decision analytic activities. The systems need to be criticised now from an empirical standpoint. For example, do they perpetuate the decision biases they are supposed to defeat. Do people use all of the data they generate in their eventual decisions, or only part or none of it. These questions remain and are the remit of a more statistically oriented enquiry. The basis of that, in my estimation, could lie in a hybrid project looking at blending subjective Bayes decision analysis with these techniques.

5.6 Discussion two: Towards a proto-Bayes

5.6.1 Summary of this section

This research stands as a kind of acceptance of the subjective Bayes paradigm. The Hurisk and Descartes systems are described as 'near Bayes' systems because the approaches can be shown to be inherently similar. A hybrid of subjective Bayes, showing the benefits package of these two systems, it is suggested would be very interesting and potentially very powerful development of this work.

This research stands also as some kind of rejection of formal subjective Bayes modelling. This is on the grounds of complexity, acceptability, time and usefulness and other compounding factors. The tension between academic and applied decision modelling is shown here not simply to exist in the tensions of rendering the subjective into objective terms.

The many tensions thrown up by more involved subjective Bayes modelling techniques are discussed, they include: complexity (industrial modelling spaces become fiercely complex); supervisory issues (an expert modeller is required and heavily involved); investment (building of a decision support model, even a subjective one, is a very lengthy process); estrangement (decision conferences and the like are very 'unreal' settings) and potential over-confidence (a tendency for people who have invested so much in them, to run models as if they were true reflections of reality). The core maths of a Bayesian approach is demonstrably hard for non-mathematicians to use in framing and communicating the uncertainty in their decision spaces.

Overcoming these challenges is the focus of subjective Bayes notion of 'requisite modelling". The benefits of this very high standard for modelling are discussed, as

are the costs. The conclusion is for Bayes to work in applied settings like those under discussion seems to require large, powerful and interesting processes to gain insight and a more coherent and consistent model of the decision world. They do so at a cost of complexity and time which users in my case studies intimate they are unwilling to pay.

The relatively inexpensive (in complexity and time) 'reasoning semantics' on offer in heuristic systems are shown to be highly valued. These are able to produce more transparency than the mathematically based semantics of a Bayesian approach. This is because they are working in a robust way and with far less complicated maths but fewer and larger assumptions. This is also a rout to relieve many of the tensions thrown up by the more involved subjective Bayes techniques.

My suggestion is that one could start from a desire for these benefits but move in a direction of 'for experts' systems. The decision maker working with for experts concepts will be required to build a narrative about their worldview and measure the elements carefully but at far less cost. This is the central technology transfer argument of my research. To get results like those shown here one needs a much more open ended approach. Also, I will argue that the resultant systems had to have the hallmark of the parent organisation, in metaphor choice, in measurement design, and in technical expedience vis a vis outputs, communications etc.

In conclusion Bayesian subjective methods do create a transition process for the decision maker to better reason about "self and world' which goes beyond optimising strategies based on expected utility. As such these approaches have a really strong resonance with the ideas behind 'for experts' systems.

Subjevtive Bayes methodology aims to mirror what real users actually think and do and then not to stray too far from that grammar.

My approach could be described therefore as a "proto-Bayesian" modelling form addressing many of the same stages and same benefits but rejecting (at this stage) any overly formal mathematical basis as the de-facto language for these processes.

To support this argument a direct comparison with subjective Bayes is made from the high level perspective and then at a deeper level of methodological intent. This exercise also gives insight into areas where future research could improve the coherence:

5.6.2 Discussion

I have argued that my work lies aligned with Bayesian decision modelling. This is not simply because of limited measurement or applied definitions of risk which are intangibly subjective, although those play a part. This is because I believe that the approaches are inherently similar and that a hybrid would be very interesting and potentially very powerful.

The research direction I chose however stands in one way as some kind of rejection of formal subjective Bayes modelling. This is on the grounds of complexity, acceptability, time and usefulness, and so on. In the next section I want to outline the precise nature of this difference in defence of my own approach. After I have done that however, I want to re-visit the idea that the systems are not so far apart as to be irreconcilable and indeed that a combination of the strengths of the two is a clear

recommendation for a future and potentially highly fruitful partnership in further research.

Positioning within a wider statistical tradition

Before discussing heuristic techniques for risk reasoning and decision support. It is important to outline an overall narrative to help position them. This is my understanding. . .

Academically it is surprising to learn that probability and uncertainty are very young techniques. The proofs and the theorem which support this kind of work are only around 50 years old. People have only been trying to apply them for 30 years, helped a great deal by the availability of inexpensive and powerful computers. This has done a great deal to boost the applicability of statistical reasoning to real world problem spaces.

In the 1930s probabalists created group theory based on the more historical probability observations of De Fenetti, Fischer and their contemporaries. Probability took on an axiomatic character, although still very firmly within an academic context. The early pioneers were not arguably very interested in whether probability was real or not even in the gaming room. It was treated like maths and therefore interesting proofs and theorem were developed. These were rendered it into the same reference grammar as addition subtraction and so on. Once that was achieved proving new theorem was regular mathematics.

Bayesians seemed to say to each other, that to be about reasoning, these axiomatic rules looked fit for purpose and so, in a distributed sense, "decided" to work within

them. They thus became the reference grammar of Bayesian modelling also. This new maths seemed to be agreed to be the simplest semantic which one could use to express uncertainty. Since the theorems had been proved, they became a formal template for deduction and so axiomatic probability rules became a set of cultural truths for objective Bayesians.

Schaffer and Dempster developed their complex 'belief theory' arguably to widen the applicability of this formal Bayes reasoning. At the trivial level this meant that rather than stick to single probabilities one could now put upper and lower bounds on things and get an interval. This is still using a controlling story of betting and betting equivalents. The semantics had improved a little because, unlike formal Bayes approaches up to that time, you were able to say "I don't know", or I am "more or less uncertain'. This was something a single elicited probability couldn't do.

Larkey and Kadane and others opened up the game theory yet further and the applicability of Bayes models within it. This remained within the traditions of the dispute over whether game theory could actually add anything to real life decision making where it proved predictably unreliable. Belief theory and game theory opened the way to more extreme forms of similar sorts of fusions of objective and subjective logic. Fuzzy logic came along, and eventually there were theorists who were just returning to expressing things as dependencies or conditional independencies.

This highly incomplete, potted evolution for these approaches is enough to show how the discipline has wrestled with two key questions.

1. Was academic versus applied, in short should the interest be about mathematics and statistics or about actual live decision making and real decision makers? That debate

was not always about objective versus subjective reasoning, often it was about the tensions of rendering the subjective into objective terms.

2. The second challenge is about the different sets of 'reasoning semantics' on offer. How is it possible to get to a set which is easy to comprehend, communicate and apply? In all cases these semantics had to be qualified by core maths. The core maths itself was arguably hard for non-mathematicians to understand. The qualified maths in reaching for an applied adaptation was arguably even harder. These techniques were stuck, to a greater or lesser degree, in this dilemma. Whilst the theoretical maths and statistics was fine, communicating a level of uncertainty between non-mathematical users in an applied domain was a struggle.

The belief theories and the game theories of this world still, I am informed, don't see a lot of applied use because, although mathematically coherent, they don't get used in real world decision settings. This failure is at what I would call the technology transfer stage.

Taking the debate forward

In these sorts of historical cases, what we have, crudely put, is a set of formal rules and then some mechanism to communicate uncertainty. So in the case of my heuristic reasoning systems, I too have tried to set up formal rules and I allow people to communicate uncertainty. In some ways I am operating a bounds of certainty argument. However, I am doing it in a robust way and with far less complicated maths and fewer, but larger, assumptions. What my approach has to commend it is that it has transferred into real world settings and if there is to be a marriage between these two thought systems, there the marriage should begin.

Aligning (or not) with subjective Bayes

Normative decision analysis seems to undergo an indigenisation process when used on real-world decisions. This seems to be the case with the transition from objective Bayes to subjective Bayes. The move out of the model-world of hypothetical mathematics and into real decision spaces causes the rise of a Bayesian apologetics. Importantly this not just a mathematical or philosophical shift, it's methodological too. Numerous profoundly psychological (counselling even), techniques are thus attached onto the normative structure. Their justification would seem to be to create and retain a sort of ethnographic form of mathematical accuracy. This can only be done if one improves the interface with humans in their non-maths reasoning.

Issues with subjective Bayes techniques

Subjective Bayes techniques would throw up a number of issues if they were to be applied to the kinds of decision spaces in this research. These issues could be summarised as complexity, supervisory issues, investment, estrangement and potential over-confidence.

Complexity: It's perhaps easier for a non mathematician (or a non financial modeller) to see that the modelling spaces in subjective Bayes decision analysis are fiercely complex for the mathematically uninitiated. The use of probability alone is obviously a concern, this is evidenced in the column space given in Bayes text books to developing methods to stop lay people becoming confused and illogical. What is evident is that this is because it is nearly universally observed that they do. This is not just in terms of key biases, a testimony themselves that probability is an expert domain, but in terms of the documented difficulty lay people have in manipulating

mathematical probability to correctly convey things which are easily understood as a narrative (e.g. dependence).

The complexity comes as de facto attribute of modelling the human physical world using truly representative numbers, formulae, distributions etc. These numbers, however sophisticated, can be likened to giving the non mathematician a phrase book to speak maths. The subjective Bayes user still has to learn to communicate in expected utilities and so on. This is a potential trap because these phrases are so hard to learn. To communicate beyond them, not to mention with them, could become exponentially more difficult with increases in computational complexity. This would be, in my terms, a key driver of technology transfer failure.

From an art form which was designed originally to be far more tentative, attribute elicitation has attracted the trappings of a more fully deterministic system. The fundamental basis of decision modelling, the elicitation of attributes, has become increasingly high powered with the advent of more and more powerful computers and the associated decision analysis software (again itself constraining and occluding in some measure what goes on in the process).

The advantage I see in building more transparent (admittedly more simple therefore) systems is that the complexity is sublimated to a few small calculations which themselves are real time updating. When the user is focussed on the "headline heuristic" through direct manipulation of a metaphor e.g. do I concur with this comparison, or does that distance equate to something I think, there is nothing confusing happening and the communication of the results is direct, immediate and natural.

Supervisory issues: To perform a subjective Bayes decision analysis the users require statistical expert support. That support is both in terms of involved analysis guidance and supervising computational activities. Short of users becoming expert themselves the modelling cannot easily happen in the absence of a trained scientist. Scanning the lists of challenges in Bayes best practice, even without the literature on biases and heuristics, demonstrates this on its own. This modelling design is the domain of a multi-skilled expert in decision modelling.

There is an argument about being bound to the utility of these models. The supervision overhead is justified when they are aimed at a singular application in a complex space e.g. plans in the event of nuclear power disaster. Suggesting that once the model is constructed the user is free to "run it" as part of decision making would again overcome reservations about the supervisory support needed to get it. Note however, this is a very restricted set of decisions indeed. Note also that "black box" approaches like this are is supposed to be counter to the subjective Bayesian's aims.

The advantage to building softer, more heuristic, models is their modular adaptability to new situations in the organisation e.g. by replacing or refining a score-card one can have a new proposition with no damage to the rest of the model. This cannot easily be claimed for a Bayes model (although Bayes Nets are perhaps built on this kind of premise)

Investment: The building of a decision support model, even a subjective one, is involved. The larger the decision space the greater this investment will be. Even small decision spaces incur a attribute and utility elicitation overhead which is

considerable if it is to be accurate and representative. This is new time and learning investment because (outside of insurance and finance) the core skills of the users are an unlikely match with Bayesian updating. The outputs, however satisficing these are of statistical standards, will be opaque to a certain degree. This will require investment in new expertise for interpretation and communication. The user needs to be able to understand and work with these outputs (a core proposition for quality control) and they will need to be a convincing communicator (or translator) to their line responsibilities. This itself may be quite challenging.

The argument for the cost benefit of such an investment in large scale energy or waste management problems is clear. In day to day safety and issue judgements, particularly those which are identified at the "contributory" level of those in the NATS project, it is not so clear. Even those very large issues in the Unilever project are very small compared to market share, profitability etc., all of which allow for highly complex modelling because of their importance. A fitness for purpose argument begins to shape itself.

Estrangement: Building decision analyses, I would argue, takes people into very rarefied places outside of their comfort zones. Although this is considered a helpful method of improving rationality, reducing bias, and focussing attention on key decision drivers and so on, the decision conference is likely to be a fundamentally "unreal" setting for many users. Sometimes up to two whole days of attribute elicitation and scoring in an atmosphere of learning and assimilating the technique at the same time, is a lot to ask. It is straightforward psychology to suggest that people come to value things which cost them a lot. If adequate learning and internalisation

does not take place (that is of course controllable but requires more of the investment above) the resultant models run the risk of functioning as black box approaches.

Models to reason with in Hurisk and Descartes have a direct utility to the input processes. The user inputs directly manipulate the core object of interest, riskiness of a future event, the likelihood that an issue will do harm and so on. The model of measurement is already in place (although investment in measurement is clearly controlled by the user at all times). This makes using the systems "direct" in the same way that subjective Bayes systems are ultimately about, but cannot provide at the outset. Hurisk and Descartes make generic and more simple kinds of decision support leaving very much more in the hands of the user.

Potential over confidence: This is not something that I am unique in pointing out, it is clear from texts on Bayesian analysis that the limitations and assumptions in a model carry forward. However, psychologically speaking, when users have a complex, expert backed, expensively invested and perhaps not fully understood system available. It would seem hard not to focus on the things it can do with some confidence.

The advantages of a decision support system like Hurisk and Descartes as described above is that they are being viewed as a useful tool, but not as something that gives the answers. A successful subjective Bayes decision analysis one must remember is aiming a lot higher than that and people would have due course to feel more confident in its output.

Some further challenges: High standards of modelling

Phillips notion of 'requisite modelling" described earlier is a key driver for a clear Bayes model. This states that there should be a cyclic process of gathering insight at the heart of building the attribute and utility models. French suggests to build a prescriptive system which is requisite one has to understand the reasoning level (consultants, consensus group or bench-markers) because the maths will differ between combination and summarisation:

"it is important to distinguish to which class of problem an particular circumstance belongs, because the means to its solution almost certainly depends on its class"

Requisite modelling adds detailed criteria for methodology choice (explicit axiomatic basis, lack of unsettling counterexamples, feasibility {of modelling and maths level}, transparency {of inputs, calculations and results}, robustness {sensitivity analyses etc}, philosophical compatibility {within the users worldview}). On top of these criteria checks of uncertainty also have to be made for input and output uncertainty (although these have a decision analytic use as well). All of this is required to help organise information into a coherent picture then utilise it to calculate the best course of action, communicate why it is optimal and provide a framework for the criticism of ideas. With all this preparation driving at clarity and validity there comes a confession, again from French:

"one form of uncertainty which neither sensitivity analysis nor uncertainty modelling can address is lack of clarity on the part of the decision maker. Often they approach an analyst lacking any clear vision of what they want or mean. Problem formulation tools can be used in this instance"

And his suggestion in this case is that soft OR methods can be applied to iron this out. In short, if your highly complex modelling does not seem to be working out, reach for another complex modelling tool to sort out your participants.

Subjective probabilities?

The assumption that subjective probabilities can be fitted to a classical model (to benefit from its axioms) is one of the cornerstones of Bayesian decision analysis. Subjective probabilities are not used without controversy. Evidence has already been cited for phenomena, such as conservatism where, in the face of clear logic individuals (using probabilities under guidance rather than as experts) still produce exclusive probability spaces which do not sum to one. To maintain consistency there are a number of fixes the decision analyst can introduce to stiffen up subjective probabilities (Using serendipitous betting schemes, breaking events into components, avoid very large or very small probabilities, differentiate joint probabilities). This again raises questions about how much and how long lived the decision support clients really need to make successful use of decision analysis. As Smith points out:

"In practice the probabilities you elicit will invariably not satisfy the probability axioms if you happen to ask enough questions, though the client is usually prepared to adjust them when you point this out and treat the inconsistencies as measurement errors"

It is a given in this statement that you (the probability expert) will be there to adjust them. This decision analysis is perilously close to being not only an art-form but necessarily the work of a master. On top of everything that organisations have to do to elicit a coherent mathematical model, they have to put systems in place to achieve

all of these other criteria as well. One would be forgiven for thinking that unless you employ a Bayesian full time, the resultant systems are going to be hugely difficult to design, use and maintain. Naturally, if you want to offset a nuclear disaster, you would be willing to have this.

In mainstream industrial decision making about small risks and medium sized issues you are this sort of investment is unlikely to be warranted. This is where a system for reasoning which bridges the formal and the sub-formal decision support systems is so necessary. Taking all of the above discussion into account and considering the ranges and kinds of decision makers that I have been working with in this thesis the key criticism of subjective Bayesian Decision Analysis has, sadly, to be clear:

It is much too involved and much too complicated.

Coming back to first principles

The heart of subjective Bayesian decision analysis is to help organise a decision makers' information into a coherent picture then utilise it to calculate the best course of action, communicate why it is optimal and provide a framework for the criticism of ideas. For me, given the distance which such an approach has moved away from normative techniques a question emerges: How much is subjective Bayesian decision analysis actually about decisions now?

It is true that at the end of the analysis decisions have to be scored. However, when one considers the modelling effort it is possible to observe which is required to make sure that these are scored, compared, and then chosen wisely, the decision itself seems almost trivial. The necessary improvements to 'objective' Bayes to make it work in applied settings seem to me to create a much larger and more interesting process whose primary aim almost might as well not be the analysis of decisions. Its purpose

could easily be the generation of insight, understanding and a more coherent and consistent model of the world within which the decision takes place. My suggestion is that one could start from these benefits thus moving in the direction of 'for experts' systems.

The reason I think subjective Bayes is like a 'for experts' system is really by its own admission. When one looks at the great many supplementary (and non mathematical) activities (however founded it is in improving or steering the maths) needed it seems to head in that direction. Admittedly the motivation for these inputs is to make the maths apply to a final decision in a credible way, but that does not cancel out these benefits. I don't think it is a controversial statement to suggest that if a subjective Bayes modeller were to quit the organisation before the utility results were in, almost everything that had been contributed thus far would remain useful. Useful in the sense that the organisation would still have had great support for planning their strategy and understanding it within their world-view.

The decision analysis user working with a subjective Bayes system, as a modelling requisite, has to build a narrative about their worldview. That relates to how this decision, and critically this decision support model, do or do not fit well with the present and future state of that. This model is ideally applied across the largest group of decision makers possible. People who share people a common set of strategies and goals and who, it is hoped, can share a common probability distribution at the end of the game. Subjective Bayes decision analysis when phrased like this looks, to me, like an organisational optimisation approach in fact it is clear that one is <u>not</u> able to choose a truly optimal decision unless it creates an atmosphere of better reasoning across the decision space.

The decision analysis user working with my systems will have a very similar experience. They are also required to build a narrative about their worldview. In my case it is a narrative about how this priority portfolio, and the heuristics which model it, represent the tension between the current and the future states of that world. In the risk and strain case this is a raft of future possibilities rolled into a collective but they are still individually available for inspection and introspection (moreso than in Bayes). In my approach the optimisation of the organisation through its portfolio choices is done by having better upstream group reasoning as the primary goal and more rational (defensible) priority narratives at the end of it. Importantly the core data is still visible and interpretable in its raw form.

5.6.3 Combining heuristic reasoning and subjective Bayes?

If I say my work is highly aligned with subjective Bayes then why did I not seek to use a Bayesian starting point? The first answer is easy. The user engagement at the concept phase was the key to my central technology transfer argument and that required a much more open ended approach. The second answer is like it, the resultant systems had to have the hallmark of the parent organisation, in metaphor choice, in measurement design, and in technical expedience vis a vis outputs, communications and so on. That wasn't going to happen if the paradigm was simply decision analysis with utility maximisation. Importantly, I didn't know at the beginning of these projects to what degree the decision support idea would get to the actual end of line decisions. These decisions were (and remain) in any case not sufficiently static and well defined to use a Bayes approach.

There were other drawbacks too. The obvious way of introducing a Bayes like paradigm is of course to attempt the issues and risk prioritisations directly from a Bayes model, or at least to do so comparatively. However, I wanted to reject of the notion of highest expected utility for a focus as (a) I was unsure it would ever sufficiently meaningful to non statisticians and (b) it was too restrictive a philosophical and methodological starting point. Taken together these are liberating arguments. Otherwise there was no need to conduct my research at all since existing mature Bayes approaches could have been applied. And in my view, would simply have failed by being too complex, too costly and too alien for reasons already discussed.

So where does the alignment come from?

Subjective Bayes modelling sits firmly between normative and descriptive decision science and, as has been argued by French, Smith and others, is itself a form of "prescriptive" decision science. What's attractive about Bayesian subjective methods to my work is that they create a transition process for the decision maker to better reason about "self and world". This clearly goes beyond optimising strategies based on expected utility. Two ideas are key first, the "best available algorithm" approach of the subjective Bayes modeller, second, the cyclical modelling of a parsimonious and acceptable (to the user) solution. I would argue that these approaches have a really strong resonance with the ideas behind 'for experts' systems.

This resonance could be brought to bear on a wider array of industrial decision making setting if only we could improve the technology transfer. As I have discussed, I do find subjective Bayes methodology inherently sympathetic to the psychology of the user first and to their mathematical requirements second.

Subjevtive Bayes methodology aims to mirror what real users actually think and do and then not to stray too far from that grammar. I have also suggested that a full power version of subjective Bayes decision modelling however, does still stray too far.

I am not the first person to propose that 'near Bayes' approaches retain some merit. The satisficing argument of Dodd et al discussed earlier goes some way to bridge naturalistic decision making theory high ground and an ex post facto Bayesian formulation. My main point from this would be that Bayes has itself tried to build these bridges. My work offers another example, from another perspective, for more heuristic forms of "good enough" Bayes modelling to be attempted.

My approach could be described therefore as a "proto-Bayesian" modelling form. I address an indigenisation of many of the stages and benefits of that approach but I reject (at this stage) any overly formal mathematical basis as the de-facto language for these processes. I prefer, in common with some modern decision theorists discussed, a narrative approach. That narrative approach does not mean that it is alien to the fundamentals processes of such as eliciting attributes, utilities, combination and comparison rules and so on.

My heuristic approach, is mediated by technology transfer i.e. acceptability. That approach is satisfied by being 'quite formal' where subjective Bayesian methods will always strive for formal. I contend that in the sectors where I am working quite formal may have the upper hand, remaining as it does more grounded within the reasoning worldview of the decision makers and their own rationality. Importantly this grounding leaves an open door to future fruitful work. If systems like mine can

achieve high levels of transfer, then an improvement on them which makes them more powerfully Bayesian is a next step.

I say this because when one looks closely at subjective Bayes it has the same output as me. Subjective Bayes requires not just a course of action (this would be 'objective') but the ability to coherently communicate the rationale for why the course of action is believed to be best. My tools are purposely designed for disputation among decision makers about their priorities and views of risk, scoring and issues over time at the upstream stage.

I have shown my participants accept that argument. They also concur that the way variables are used and the range of formats in which they can be compared is a more focussed and more rational platform for argument than the previous cultures of unstructured lobbying. These accepted reasoning platforms, such as the comparisons of decision maker emphases, strain weighting profiles or 'now and future' predictions for issue importance, are inherently less biased than board-room style debates they supplant. The focus produced is perceived as refreshing, efficient and effective by users as evidenced in its acceptance and deployment.

5.6.4 Direct comparison with subjective Bayes approach

The tables below show how the systems compare with their subjective Bayes from the high level perspective and then at a deeper level of methodological intent. This also gives insight into areas where future research could improve the coherence:

Table 19: High level comparison with subjective Bayes

Hurisk and Descartes	Subjective Bayes
At the high level comparison	
Allow users to reason with their data on a number of levels within and between issues	Developing and understanding of the needs for, and form of, expected utility
Elicit an idea of the priority of issues and risks now and in the future	Elicitation of probabilities
Provide solid first wave evidence to reduce the number of priority issues and risks to a manageable amount and communicate those	Attribute elicitation
Collect data from all relevant experts on base set of risks and issues	Attribute elicitation
Provide evidence to brief the board project teams on top priority issues	Performing expected utility calculation

Table 20: Methodological comparison with subjective Bayes

Hurisk and Descartes	Subjective Bayes
At the detailed methodological level comparison	
Chief to the design of both systems was to deliberately blend the 'hard requirements' of good measurement and strong definition for an issues and risks prioritisation (and process) with the 'soft needs' of key stakeholders (be this look and feel issues, measurement scale choice or output).	This approach is analogous to the elicitation of attributes phase in a subjective Bayes model. A great deal of effort is invested in understanding the user's attribute models from within their world-view. This is in the understanding that without this the results of the utility maximisation (which is only an incomplete mathematical rendering of key aspects of that world) would be meaningless.
The development process of both tools included the measurement of attributes, the comparative techniques and the reporting elements of a suitable tool always being tested in the fires of acceptability for use in the live setting. This was done with key user groups. The resultant iteration (often simplification) of their design was key.	Once a model has been built there is a great deal of further effort in configuring, adjusting and refining both the 'requirements' and appropriate measures that might reflect them. The use of formal sensitivity, calibration and reliability measures back on the mathematical profile is a formal step.
The deployment of Hurisk and Descartes at the portfolio prioritisation level, rather than the individual decision point, was a key part of the design. This creates an overview on multiple levels and allows "comparison of apples and oranges" on a common score-card. This was an essential user requirement. The resultant risk and global issues prioritisation were portfolio led which meant that they could communicate across the company.	Subjective Bayes, at a lower level is a portfolio technique. The overall utility score is essentially for different portfolios of options. In terms of comparing incomparable things it is quite common practice to simply ignore uncertainty at this level and just give different portfolios a numerical score and find the example with the biggest score. This ignores uncertainty associated with the scoring methodology.
Prioritisation exercise results sharing through formal reports and logical presentations, often followed the narrative of the analysis itself e.g. tracing the identity of a strategic priority from its group scores at the technical level. This meant that even at the "recommendations to the board' phase of simplification, the reasoning audit trail for the entire portfolio analysis is available.	Explanation and audit trail are central to subjective Bayes on two fronts. First, for the legitimacy (to the modeller) of the expected utility maximisaiton led choices. Second, mindful of the limits of users to blindly accept a mathematical priority, coupled to their obvious communication needs in selling on the results in their businesses, explicit reasoning, questioning and over-riding are central to subjective Bayes model resolution.
The process for issues prioritisation had to be formalised within the very dynamic developments of the "NATS" re-structuring and the "one Unilever Issues Management process". This responded in both cases to radical new operating frameworks. In Hurisk and Descartes this was a generic background force (often represented at the micro level in risks, strain questionnaire items etc.)	The use of decision conferencing as a conduit methodology for subjective Bayes techniques recognises at the outset that formalising the influence of strategy limitations and available choices is a determining factor for a successful system. This is a formal (resource intensive) technique and a success criteria for this kind of modelling.
Scoring within Hurisk and Descartes is in a formal hierarchy. This is to allow users to select what level they wish to score at in terms of time and simplicity, however the net effect of scoring at the detailed level is a roll up of multiple scores into two high level attributes which are of course then plotted.	A key technique in utility maximising models is the use of the influence diagram where multiple weighted attributes lie at the bottom of an attribute tree and are rolled up through an algorithm to create one overarching utility score which is contributed to by the scores underneath it in the attribute structure

5. 7 Discussion on possible advantages of a heuristic approach

5.7.1 Summary of this section

Research into, and application of, the Hurisk and Descartes systems has expanded and applied the idea of heuristic reasoning to create user-centred decision support tools. Both tools are real-time reasoning systems which can equally support live and off line group reasoning. The notion that these tools represent a "heuristic reasoning platform" is investigated and this is described as: "a device which imparts the power to discover through rules of thumb using incomplete algorithms computationally, visually and in terms of process."

In the consideration of the tool design, phenomena which are classed as "essential natural heuristics" are explored and discussed. These heuristics enable people to reason about, and creating summary classifications for, complex realities in a simplified thought matrix. Some problems of reasoning associated with them are discussed.

The advantages of including improved versions of these reasoning modes in any decision support system are explained from the point of view of technology acceptance. The way in which these tools can be used to carry a package of improvements to reasoning is therefore considered a highly effective use of them. This can be done by working within the constraints which they bring and also by expanding the heuristics themselves to suit accepted organisational decision support needs.

The use of these simple tools is seen to address the large overheads associated with more formal methods and this trade off is suggested to come out in favour of heuristic

approaches and their real and perceived benefits. Three methods of using existing natural heuristics of decision maker groups and developing reasoning tools from them are explored. These cover visualisation techniques, score-card development and the overall intellectual housing for these approaches in constructed concepts like "priority" and "risk". How each of these elements conforms to a definition of being heuristic is explored in detail and aspects of the resultant functionality of the two systems are described.

5.7.2 Introduction

The focus of this thesis does inexorably surround the applications of a primarily philosophical notion. That notion is how something is a heuristic reasoning tool rather than a deterministic reasoning tool. In the following section I have expanded on my understanding of how this applies to the Hurisk and Descartes tools.

In developing tools to support expert judgement in air traffic control risk and FMCG issues prioritisation I have shown how to expand and apply the idea of heuristic reasoning. This has been done to create a user-centred decision support. In this section I'd like to take a more detailed look at the two systems. The reason for this in a stand alone section is that the systems are not trivial pieces of work, it will help to read about them in a concentrated way.

Both Hurisk and Descartes are real-time reasoning systems. They can be used across internet meetings, they can support group reasoning and they can be used for off-line individual reasoning with or without a view to some future collaboration. Users agree that these improve their abilities to consider complex propositions and come to decisions about them. A lot of the elements are like heuristic versions of what would be needed for formal Bayes systems.

Why Heuristics?

I propose that decision support comes in improvements to users own heuristic reasoning around complex and uncertain entities like risks and issues. This is predicated upon a particular usage of the term "heuristic" itself. Much scholarly work has resulted, as we have already discussed, in a sub-field of decision science called "heuristics". However, that specific scientific application cannot be allowed to

commandeer the whole arena. There is much to be gained in taking the term back to its roots.

Heuristic comes from the Greek "to discover" an it is at hear a philosophical construct with many applications and definitions. There are two that interest me most. First, from the Psychology of thinking, a heuristic is a "rule of thumb". That is a thinking process known to be limited but useful in real time because of its speed and economy. Second, from computer sciences, heuristic is seen as "reasoning in the absence of a complete algorithm". A fusion of these two ideas suits best what I have tried to achieve in this work. That is what I connote by heuristic and this use of the is implicated in the two heuristic reasoning platforms I have designed. This idea might therefore be described thus:

"Heuristic reasoning platform". A device (not restricted to software) which imparts the power to discover through rules of thumb using incomplete algorithms computationally, visually and in terms of process.

I further suggest that certain phenomena could be classed as "essential natural heuristics". These are observable in decision maker groups and I have recorded some of these. Simple examples are "high, medium, low" approaches to more complex objects; vague (and flexible) rank order scaling tasks, such as post-it pad exercises; loose categorisation of "co-ordinate" like objects into "boxes" e.g. "key players", "Cat 3 risk"; and so on. All of this I would classify as heuristic in form. People are reasoning about, and creating summary classifications for, complex realities in a simplified thought matrix.

This reasoning is doubly heuristic because it is subjectively referenced (high to you might be medium to me and so on) and no checks are put in place to calibrate, mathematically or otherwise, what is going on between and within users and uses separated in time. I have observed in field study that there is usually no time for such activity. Even if there was I have not observed an appetite for it.

So the heuristics which interest me exist "in the wild" so to speak. Groups of decision makers under time pressure are using these as labour saving devices. These individual are looking at abstract externalities and expressing them in numbers, colours, positions etc. These expressions themselves are of course borrowed images, from maths, list writing, even from road signs. That classes the whole area to me as "fundamentally heuristic".

Problems arise in the organic use of these techniques of course: inaccuracy and lack of repeatability; social hierarchy effects (e.g. no-one wants to contradict the boss), averaging effects (taking a vote), loss of rationale (someone wipes the white-board) and a host of other drawbacks. However, these heuristics are, I would argue, "natural". What I mean by that is a lot like the sense in which a plant might be "naturally occurring". They just seem to be there. This means that people use them, like them and may consider them in some way to be a professional skill.

Since that is the case in the design of decision support systems one cannot just dump them. The more critical the decision they are applied to, e.g. threats to competitiveness, or preventing breakdowns in safety that could lead to loss of life, the more important making them better becomes. There are, for the purposes of this thesis two ways to make them better. One is replace them e.g. introduce decision dialogue conferencing, Bayesian decision analysis etc. The second is to work with

them but to change them to be more like something that would satisfy such a decision dialogue process. Having chosen the latter route the challenge was to try to do this maintaining the "integrity" of what these things already do i.e. make people feel useful and listened to, allow people to display their expertise, be manageable in a heavy schedule and so on.

This was being done for an important scientific reason however. That reason was to subvert that very benefits package by introducing higher order benefits, increased rationality, better accountability, clearer communication. These are being deliberately introduced through methods which only ostensibly to "support" the existing approaches. I think the case studies in this thesis have shown that to be a very powerful approach. This is my idea of expanding the natural heuristics from within.

I don't dispute the usefulness and value of formal decision science at all. However, it has key drawbacks which can be described in a number of ways and it is possible that decision science practitioners will agree less or more with these. The mechanics of introducing expert decision sciences into live applied settings is very high impact in at least three ways:

1. Overhead: The organising overhead is often huge. This has to create a local capability to carry the weight of learning a new methodology. I know from field experience, that is a barrier to success.
2. Complexity: It is normally introducing a completely new expertise to the users and usually a mathematical one at that. A barrier in non mathematically dominated areas.

3. World-view: It imprints the decision making with its own world-view. I would say this is true even of techniques which try to encapsulate the users world-view in their execution.

If one were to need evidence for this third point then dropping back to the decision science definition of 'heuristic' provides it. It is very clear that the prevailing worldview here is still that human beings are severely limited as rational agents. Their capacity for rationality is considered low unless they have been formally trained (by decision science gatekeepers) and make use of powerful external computations (this is of course notwithstanding upcoming counter-views already discussed).

These things spell for a psychologist "technology transfer failure" or at least limited transfer. When this is compared to the kinds of, sometimes nonsensical, ideas that shoot up rapidly in management science this raises a question. Why do these, sometimes highly over-simplified, ideas seem to have a higher penetration in organisations and a longer half life?

My thesis is that we have to meet experts (not decision experts) where they are with the tools they are using and lead them on a journey to improve these. The journey has to end in a demonstrable benefits package which the experts have to venerate and own. In the case studies in this thesis I have shown high levels of success at doing this. In the next section I want to isolate some of these ideas.

5.7.3 What Heuristics Have Proved Worthwhile?

I have stressed that I would like to see natural heuristics of the decision makers and general heuristics of the decision landscapes broadened and strengthened. This is as a

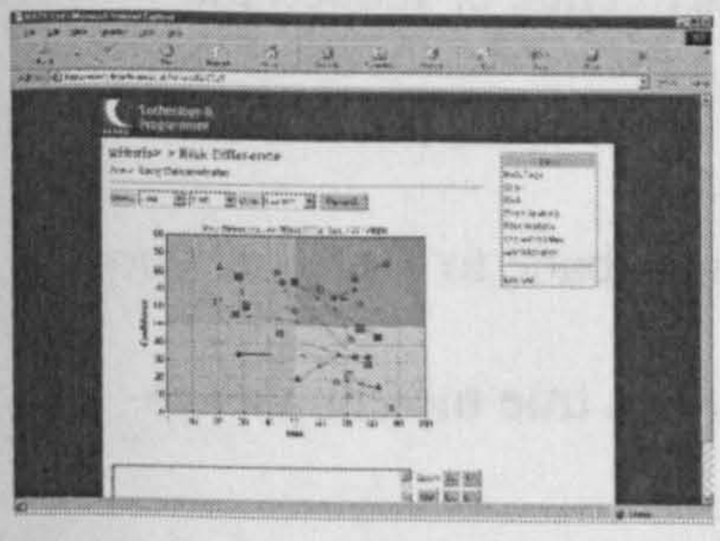
route to improving rationality whilst keeping decision support within the "natural reference grammar" of expert groups. I have suggested that such a process should result in "improved heuristics" and I have presented evidence that suggests this has been achieved. Using the dual definition of rule of thumb and reasoning in the absence of a complete algorithm, I have created tangible heuristic reasoning devices which are not just a collection of pre-existing ideas. Let's illustrate that by examples in three classes:

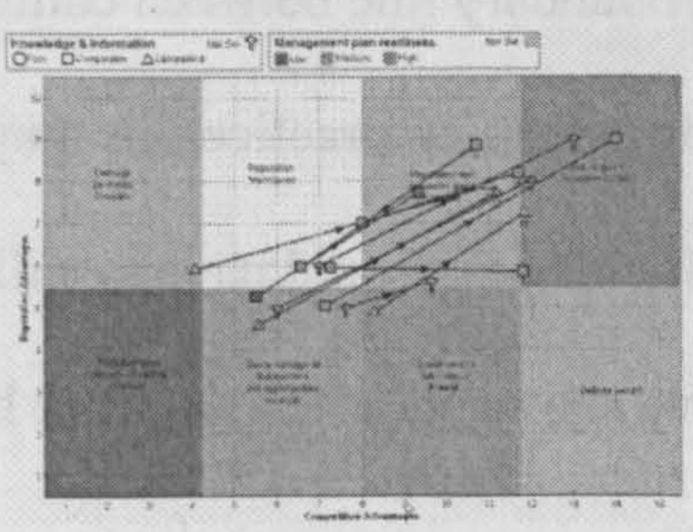
1. The visualised representation: the various plots
2. The summarised scale: the various multi-attribute scoring systems
3. The abstract construct: risk, strain and priority.

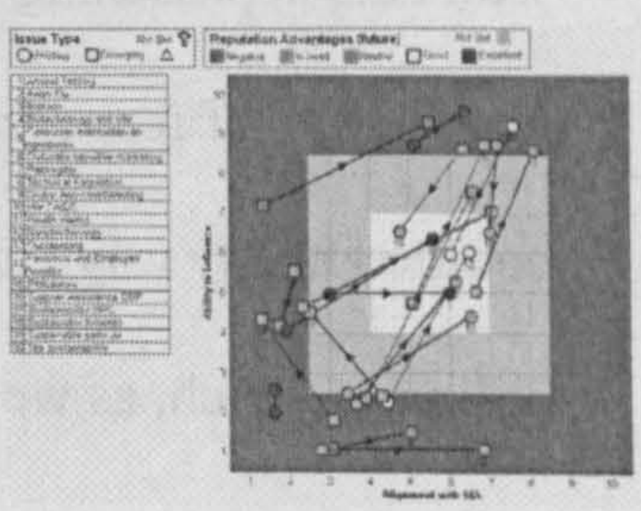
The visualised representation

The standard two dimensional Cartesian plot has an 'x' and 'y' axis and an equal scale on each. Co-ordinates dictate the position of a point, functions dictate the shape of a curve and so on. Variations involve mixing scales such as time on one axis and money on another. In its proper form it is a representation of ratio level data.

Both in the NATS project and in the Unilever project Cartesian plots are used and some of these are shown:







The reason these graphs are heuristic is this, their substance has a communication and reasoning purpose only. Any mathematical properties of the scales is at best uncertain. The measurement items are not fungible e.g. "competitive advantage" cannot have real quantities it is a true "psychometric". The scales are, at best, ordinal level data, the increment is not fixed by any mathematical function and is entirely relative to a subjective reference point. The use of a Cartesian plot to display this is a contrivance.

This is a convenient and powerful contrivance however, based on a positivist image with a borrowed authority from the world of maths. These graphs relate judgements, not measurements. From a communication perspective, they are highly expedient in comparing judgement data. This is precisely because people are familiar with the real thing from school and college and can relate to them. They have soft-wired "rules of thumb" about how to interpret what they are saying.

A further contrivance is the sub-division of the plot into a number of more or less equal sized areas. These are used to create higher-level meaning, or some might say to focussed attention. Little thought is given however to the uncertainty associated with data approaching the boundary line between compartments, or indeed the intersection of four compartments. In practice this may be noticed, but it is not codified in any way. To do this one would have to collapse back to the "pure" co-ordinates which, as we have shown, don't in any case have a true meaning as co-ordinates.

What makes these plots heuristic judgements?

These plots are therefore a heuristic formulation of real Cartesian co-ordinates plots. The areas or boundaries within them achieve for the user is not the representation of real data. It is cognitive support for a shared narrative. These visualisations are the vocabulary a sort of shared cognitive short-hand. Once again the interpretation of plot points is not about the shaky maths which put them there, it is about comparing the relative narratives their positions propose.

Seeing these sorts of plots in this way is immensely liberating because to improve the heuristic reasoning your starting point is not accuracy, it's story. Better plots will tell a better story. Improvements in their decision support utility will be borne out of looking at what they are not saying. This is precisely how the introduction of time (the delta value improvement), the addition of scrolling qualitative data, the capability to chunk data together, park it completely or the audit of the history of a point across analyses came into being. These were all driven by the stories people wanted to tell and the system evolved as people, enabled by its standard functionality, thought up their own better heuristic. I helped of course.

The summarised scale

Almost all the variables collected in these case studies were on rating scales. These scales are psychometrics. They have been derived using research methods such as depth interviewing or analysis of pre-existing data, such as strategy documents. The Unilever case is the easiest to discuss in terms of how the summarised scale was a heuristic measurement device. Four main score-cards were developed, these were:

- Competitive advantages
- Reputation advantages

- Strategy Into Action fit
- Unilever influence

These score-cards all illustrate the essential natural heuristics of the decision makers which were expanded for more powerful (meaningful) heuristic reasoning. This was achieved using the following principles:

- Use of behaviourally anchored ratings
- Absence of any numerical indicator at the point of scoring
- Direct comparison of present and future proposition on the same scale
- Hierarchical design which allowed users to move between levels of scoring (with associated propagation)
- Memory function which preserved rationality when the scorer over-wrote detailed score profiles with simple ones (as well as indicated prior scores)

Behaviourally anchored: A typical item in a management survey tool will favour a five point scale approach. This will generally vary across some generic bi-polar variable from "strongly something", e.g. strongly agree, to strongly something opposite e.g. strongly disagree. The terms at either end of the scale are called anchors. In a behaviourally anchored rating scale case the anchors relate to real behaviours which can be verified. For example in "reputation advantages" scale the negative end of the Media Behaviour scale reads:

"Company heavily and repeatedly criticised in a broad and enduring attack in a key range of media".

No numerical indicator: Experimentally scales in early case studies did not contain any numbers. It was expected that this would be unpopular. It wasn't. All scales in

the Unilever project adopted it. In the final interface the sliding line did change colour from red (very negative end of scale) through amber (mid point) to green (very positive end). This was another low-key example of expanding a general heuristic (the traffic light approach).

Direct comparison: This tool is all about stylised comparisons between narratives. This is the basis for a reasoning platform argument. The direct comparison of a "now" and a "future' narrative was reflected in the scales. The proposition for the future score was simply "if we manage this really well in the future". For each individual scale the 'now and the future' were compared directly at the point of scoring and remained in view thereafter. The length of time between the now and future scores varied within and between analyses and was frequently not set at all. Users were happy with a nebulous ideal of a future state.

Hierarchical question sets: The scales in both Unilever and NATS case studies always resolved to two variables which were then plotted on an 'XY' graph. Users were offered a hierarchical scoring choice. Direct manipulation of a data point on the plot was the most basic. Next, users could score the same two variables, this time with the benefit of the anchors (simple scoring). Finally an "intermediate proposition", which contained up to ten scale items per variable could be used. In all cases a scoring action in one mode affected the scores shown in all others.

The intermediate scoring followed by a simple scoring makes the most informative description. In the intermediate case users will provide a range of scores on detailed variables. These are rolled up to the top level. If a user were then at a different time to re-score using the simple variables options the subtlety of that detailed argument

would in a simpler case be over-written by some sort of mean score. To avoid this loss of data the underlying scale scores were coded in software to preserve their relative shape as much as was possible.

Memory function: "Memory triangles" small point indicators on the scales always recorded the position of the last intermediate scoring round. Thus if the data was over-written it was possible to retain, until the next detailed scoring round, both sets of scores.

What makes these scales heuristic judgements?

The power of behaviourally anchoring scales rather than just using bi-polar markers is in the rationality. Where the scales are going to be used in plenary you avoid the 'one woman's high is another man's medium' fallacy. People have to rate and, if necessary, defend an exact agreed behaviour, not a subjectively referenced variable.

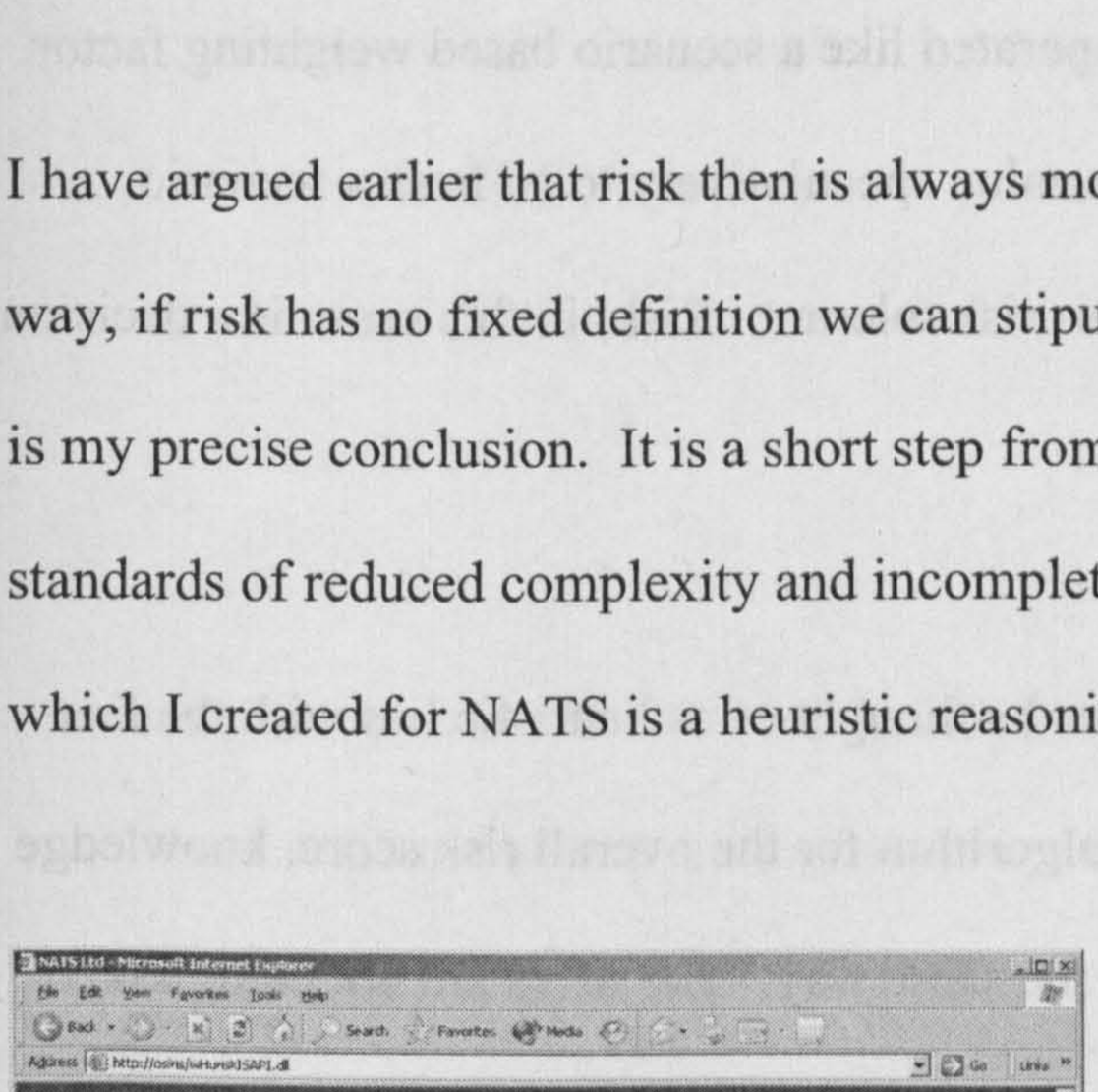
These judgements were still elicited from people however, as subjectively referenced beliefs or preferences. Although people are being asked for data this "measured world" does not equate to a "formal model world" in the scientific. It does equate to a model in the social science sense however. Scaling techniques, like those here, reduce the cognitive complexity and uncertainty associated with highly complex measurement propositions. This they do with an associated loss in accuracy. Like the plots which represent the outcomes of this cognitive activity therefore, the scaling concepts here are over-simplifications of the world and the measurement components are introduced in the absence of a complete(ly controlling) algorithm.

The Abstract Concepts

Risk

I have argued, hopefully without reaching for too post-modern a philosophical position, that concepts like probability and risk in this applied area are abstract collectives. Each of these, in the many arena where they are discussed has no definitive position. Rather the concepts reference a complex of disputed ideas. There is no doubt that they are very powerful ideas, but there is similarly no doubt that they are constructed.

I have argued earlier that risk then is always model-based ideal. To put that another way, if risk has no fixed definition we can stipulate one. In the NATS case study, this is my precise conclusion. It is a short step from there to argue, using our same two standards of reduced complexity and incomplete algorithm, that the risk construct which I created for NATS is a heuristic reasoning device.



The eight component risk model made good sense to the users in NATS. Scores on these were created using the "clock faced dials" metaphor shown. Three

measurements made up risk, three local algorithms introduced uncertainty to those measurements. These six combined to make up two summary variables which combined via a global uncertainty algorithm to create a final risk score

How was this concept heuristic?

There are three key levels to the heuristic nature of this concept. First and easiest to describe is the notion of a changeable risk tolerance score. This score operated an algorithm which altered the absolute value for a users risk depending on how they were feeling at a given point. It also operated like a scenario based weighting factor. This is because users could change the value speculatively to address a scenario within which they may feel more or less risk tolerant. Risk, in this sense is a decision support idea, not a measurable object.

The second reason why this concept can be thought of as heuristic lies with the fact that just as risk tolerance is a cardinal algorithm for the overall risk score, knowledge mediates the local individual scores in the same way. The users are not being asked to measure risk in and of itself. User are having probability, control and effect elicited from them for a particular state of the world and then arbitrated by how much they personally believe they know about that world in this case. The object of risk scoring, as is the case in scoring issues in the Unilever case, is that this measurement technique requires the user to explore and expose their own reasoning as well as provide their elicited value.

The third reason has already been discussed. An algorithm which combines knowledge and elicited judgements arbitrated by a changeable emotive component is very close to the idea of a mathematical rendition of a cognition. The users are laying

out their thoughts, and the potential variance within those, and over time, to create a state value for a risk model. The model itself is not "risk" and neither is the value. It is a conceptualisation of judgements and feelings about a shared definition of risk. In a more complex way this risk concept is simplifying a complicated measurement process to create a judgement based system which is faster and more simple.

Likewise the reasoning that people are committing to in this case is in the absence of a complete algorithm in two ways. First, there cannot be a complete algorithm for a social construct. Second, the users are not aware of the properties of the algorithms in the Hurisk tool. This satisfies our two sided argument for a heuristic reasoning device.

How is strain heuristic?

Strain is really another form of risk since the paradigm is still threat to safe operation. In the system it acts as a "sister" to risk, which is why they are displayed together at the high level summary. The incompleteness of the algorithm controlling strain scores reasoning goes without saying a 111 item psychometric questionnaire is a complex matrix of ideas to form an algorithm with. What it does have is a combination rule for the scores. What the combination rule does therefore is describe the world in a shorter story than necessary. The scores are shorthand narrative blocks, and that, in other words, is a heuristic reasoning form.

Like the risk case it is the reasoning itself which is the value of doing the scoring, because the end score is virtually meaningless without the narrative which gave rise to it. In these and many other examples within these systems what is happening over and over is that measurements are made, metrics are applied and the results are combined, usually visually into a complex heuristic. Algorithms replace real

mathematical models, judgements replace the real measurements and narrative analyses replace the real long-hand narrative that is being conceptualised. That fulfils my idea of being heuristic. It raises the philosophical possibility that other decision modelling, even mathematically more coherent modelling, might be more than a bit heuristic too.

5.8 Concluding argument

In the applied world in question in this thesis we have a condition of high-volume, small-scale rapid-turnover group risk reasoning. The theoretically very formal and heavy approaches from normative techniques would tend to de-construct and criticise the reasoning forms and point out their errors. The methodologically very heavy approaches from real world decision making would tend to build very large superstructures around the reasoning agents and, in so doing, potentially increase the available data and disagreement.

Starting from a "content led" approach however, what people have to hand and what people do already think and do to reason about risk, and improving that is arguably a highly legitimate approach. This is especially the case if what people are already doing can be made more elegant in terms of exposing available data and disagreements in a focussed way. This is especially the case if what people are already doing could be put onto a more sound, transparent and auditable reasoning platform.

I have harnessed risk, and decisions about it, in a way which those who have a high stake in these terms recognise and approve of. These people are not theoretical decision makers in a laboratory experiment on probability (valuable as this essential work is to understanding how humans do reason), they are risk users. The formulations of risk and decisions in these studies have been wrought through operational partnership and applied social science.

These people recognise the risks and the decisions at the end of this process as their own and therefore seem happy to drive the quality of the operational application of

these concepts. The decision support systems I have developed are not costly, financially or in terms of operational disruption. Most importantly they do not cost a lot in terms of intellectual transfer into the live setting. The systems are recognised for their contribution to more efficient process, more effective reasoning and creating better results than the existing methods and tools had done.

This work has, I believe, spanned, comfortably or otherwise, the gap between normative and descriptive decision science. The maths it uses is good maths and makes sense, but it is also elegant and fit for user-defined purposes. To develop an effective definitions base for this maths we have looked very carefully at its psychological meaning to real decision makers. This has taken me into the area of heuristic reasoning.

I have expanded the existing concepts of heuristic reasoning and taken them back to something more like their original philosophical roots. These expanded notions of heuristic have then been applied to decision support systems. These have been tested and proved to have value to the host organisations. I have also developed novel forms of heuristics for applied reasoning and have contributed to the theory in this area therefore.

I have suggested that heuristics might be worthwhile way to conceptualise a more general class of approaches and tools which one sees working in the industrial world. Irrespective of origin, or legitimacy, there are a wide range of general natural heuristics which have high resonance with the way people are used to working and want to work. Heuristics of this kind can have a huge influence on the way that people reason. Therefore they need to be included in any model, if only to take

advantage of their high perceived validity as a transport mechanisms for better versions and thus better reasoning.

I propose that an observation science which looks at what people are thinking, feeling and doing is a key starting point. Providing reflection on these is a tool to support these decision makers through augmentation of their existing expertise in order to promote effective technology transfer. I propose that the resultant 'for experts' system transfers better than an, arguably intellectually superior, expert system. These often fail to be transferred other than in highly risky, complex and multi-stakeholder processes, which are rare.

I have contributed meaningfully to understanding how people in applied safety, and business critical, settings can see and work with risk concepts. This has been done whilst respecting the idea that these people are subject matter experts, not risk experts. They look for tools to support their work rather than risk or decision science expertise. They do not, in my observations, feel the need to use the 'model completeness' standards which come from those exacting disciplines.

New definitions and measurements for risk have been designed, accepted and valued in formal operational contexts. The challenge is now to improve them. My methodology, based as it is on a social science approach whilst attempting to utilise mathematical benefits from consistent measurement and meaningful use of algorithms, is pretty close to subjective Bayesian thinking. To me however, it is a more proto-Bayes because it seeks to be more applied and be a servant of experts' own worldviews. I'd argue, that to do a really thorough subjective Bayes analysis of

risks and decisions it would, in fact, be better to start from a position which is a lot more like the one I have taken.

The support offered by prescriptive decision science can be offered more effectively, in my view, from within the decision reference grammar of an organisation. This stands in contrast to teaching them a new one. In the rarefied setting of building a Bayes model this new grammar might be accepted, but I believe it will (and does) struggle outside of that. Utilising the technologies and the languages already available in an organisation (their natural heuristics) makes a lot more sense to me because these have a track record of being considered useful. Importantly they are already accepted. I'd suggest this is more than an opinion, this research has proved the case, in two concrete examples of heuristic risk and decision support phrased into 'for experts' systems.

I'd suggest that what has been achieved here is similar to what the forms of subjective Bayes approaches I have reviewed are trying to achieve. I'd go further and insist that the two approaches could be developed into a mutually beneficial hybrid. My systems could improve in their mathematic consistency and evidence base for long term insights particularly into the way that people are actually using them to reason. Subjective Bayes thinking could benefit from the way these systems have managed to have a high penetration into the host organisations and become established and valued working practices. The Bayes approach would also benefit from being placed into an arena which is about more small-scale technical risks and ordinary work a day decision making about risk, where I think it has a great deal to contribute, rather than being confined to complex modelling environments for large decisions.

A hybrid of my approach to definition and reasoning and subjective Bayes analytical power might generate that elusive set of evidentially "fast and effective heuristics" that this area is in need of. Together these could further shape the world of applied risk reasoning. That reasoning could better meet the challenges in the extant research by blending the complimentary power bases of applied psychology (of the kind I recommend) and applied mathematics. This could produce accepted and proven effective 'for experts' systems which are deployed, valued and, above all, used.

5.9 Future research direction

The tools introduced in this research have two things in common, they are highly detailed pieces and they were both developed in a live setting under the supervision of sponsors who wished to direct the outcome and application. The resultant study is highly descriptive in nature and leaves open therefore a number of interesting research avenues which have not been explored. These are in four categories:

More formal statistical treatment of the effects

In the Unilever case study the following effect was observed. The introduction of a computer tool which delays group interaction via a process which causes people to reason their arguments in advance is considered highly desirable by the busy target audience.

In academic terms however, such evidence would convey an important face validity to the system in question. A desirable comparison exercise still to be done would be to look at actual construct validity. The effects of the use of the tool as outlined compared to a control condition wherein participants meet and discuss risks in their usual modality would achieve this. The balance of individual and group contributions to the final risk profile would be extremely informative as a differentiation of the effects of any particular elements of the tool and their interactions. This would give a more detailed analysis of the benefits and problems of the on-line and off-line modes for this group activity.

A deeper psychological assessment of the value of the effects

In the NATS case study the value of the risk assessment approach was differentiated across three target audiences using a small exercise looking at opinion data and

subjective weighting structures for a similar problem. The results of this work suggested that these groups viewed the functionality and reliability of the tools very differently.

The air traffic controllers valued the systematic rigour of approach and the blanket assessment of risk and strains as being equally valid. This seemed to conform to their operational training to do with the systematic reduction of all uncertainties. They were noted to be particularly interested in the system's ability to convey this complexity back to a management who could be viewed as over simplifying.

The managers in the study, conversely, tended to focus on the systems ability to "roll up" ideas into larger summaries. The possibility that this system could be used as a replacement for (tiresome) brainstorming exercises was seen as a key feature. Very little in depth engagement with the tools was noted in the interviews.

The engineers in NATS were the focus of a much more detailed study of the use of the tool. At the end of the trials they adopted the prototype risk assessment as an operational tool and rejected their existing (less rigorous) approach. The features of the system which clearly attracted this group were to be found in its ability to compare views with peers and in its exactness.

Further work into the effects of these different perceptions of the tool would be insightful in relation to its effect on risk assessment and mitigation. This work could explain the value to different staff groups of the scoring and recording of reasoning, and the effect this has on confidence in recommendations. From a statistical point of view, it would be useful to have an analysis of whether it is possible to attribute

perceived improvements in risk-assessment unambiguously to the tool or its effect on the balance of opinion which might conceivably be arrived at by a different method.

An explanation of tool acceptance (as a function of effect)

This study places a lot of emphasis on technology transfer. The acceptance of the heuristic approaches to reasoning are seen as a key finding. This is especially when compared to the lesser levels of acceptance of more involved formal methods.

The adoption of a risk assessment tool and management approval of it does not necessarily connote that the tool has improved risk assessment. In cases where very poor risk reasoning was observed such a tool is a clear step change and this has been discussed. Where things are less clear is in a more detailed analysis of how we might judge the tool to have improved risk-assessment.

Where possible this further research should be under controlled conditions with reference to standard statistical techniques for establishing causal effects. Candidate variables for such an examination are: testing who is using the tools, what the alternatives were, and tracing the decisions, in order to see whether decisions were made more quickly, supported by better reasoning, and effective in ensuring that relevant, rather than irrelevant contributions were elicited from users.

The improvement of the statistical elements of the tool

Throughout the discussion a clear parallel between the heuristic approaches and subjective Bayes decision modelling is drawn. In essence there is an attempt to suggest that heuristic reasoning is to subject Bayes as subjective Bayes is to formal Bayes. That is to say that the process of accommodation needed to adapt formal

Bayes for more applied settings could be extrapolated to those settings where even subjective Bayes is prohibitively complex or time consuming.

An investigation of the way in which the very powerful mathematical conditioning of the subjective Bayes approach could somehow be accommodated in the highly transferable skin of heuristic approaches is a hugely attractive research area.

Appendix one: Unilever comment

Issues Management and Issue Prioritisation – A Unilever Communications perspective

I have worked with Jon Arthur for over 3 years and, most notably and appropriately for this text, we have worked together on a project to prioritise Unilever's global issues.

When I met Jon in 2004, we were in the process of bringing Issues Management together, at a global level, in Unilever. We were standardising the way we managed issues, individually, and how we shared information and position statements throughout the business. We also had ambitions to improve our overall governance of key global themes in order to find ways to improve our response to public policy. We wanted to make sure that the senior decision-makers in the business were aware of the issues that have greatest business value – and that they were more involved with them. We also wanted them to be kept updated on the status of these issues and to champion issues through important stages, whenever needed. Jon had applied his stakeholder and risk models to smaller projects in Europe and it was clear at a very early stage that he had a skill-set which could be applied to our global requirements and the appetite for a much greater challenge.

In the early stages, Jon consulted widely and developed a scoring system to set priorities for issues. He focused on speaking to practitioners and tried to match their skills with the brief from senior stakeholders, which was, to be fair, an evolving one.

Together we made issues prioritisation one of the pillars of global issues management.

One of the important aspects developed in the early stages was the dual focus on present and future. This meant that although we could find ourselves in a state of disarray on a particular issue at the current time the tool still allowed our experts to justify the possibility of managing an issue to the business' benefit in the future. This in itself justified the need to put some resources behind that issue. In contrast, this also allowed us to identify that there are some issues where even if we manage them really well, we're not going to be able to deliver any significant business benefit in the future. This provided a step-change in the way we considered issues.

By the time the Unilever Issues Group had assumed the governance of global issues for the business, they were ready to use Jon's methodology to evaluate issues with a more strategic mindset. This meant that the Unilever global issues portfolio was shaped by the key elements of business strategy, (e.g. implications on our supply chain, implications for marketing products in key areas, implications on consumer confidence). - considerations on which Unilever Issues Group had a clearer insight than individual Global Issue Leaders.

Through the first issue prioritisation process in 2006, which was initially designed as a pilot, Unilever narrowed an initial portfolio of over 70 issues

down to around 25 high priority issues. This was then narrowed through discussion to 11 specific key issues and presented the results to the Unilever Executive and Board. They endorsed the results in their entirety and since that time, the model and Jon's expertise have been used to support decision-making and resource-setting at European and Americas regional level, and in many countries, including the UK – where the 2008 Public Affairs plan is based on Descartes analysis.

At a practical level, the key benefit of the system Jon has developed for issues priortisation is that it takes away the lobbying and horse-trading of the 'post-it' on the wall group exercises. It allows recordable measurement – either individually or in plenary – against the most important business criteria and then gives those in the business with the best overview an opportunity to discuss how to manage a well-designed high priority set. It actively encourages thinking about issues in a systematic, rather than political, way and drives good definition of the issue and how it impacts your business.

The success of Descartes in Unilever Issues Management is down to the potent combination of the model and its architect. Jon deserves enormous credit for what he has achieved in driving through a unique system through times of both high expectation and good support and some periods where there were doubts about the value of the programme against competing priorities. I am confident that in the years to come - because it is over that range of timescale that the value of our global issues strategy will be borne out – the introduction and utilisation of Jon's Descartes model will be seen as a key factor in Unilever's successful management of key global issues.

G. J. Gordon (11 December 2007)

References

Action Aid International. Power Hungry: Six reasons to regulate global food corporations. Action Aid International. 2005

Adams, J. (1995). Risk. UCL Press, London.

Ajzen, I. From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckman (Eds.), Action-control: From cognition to behavior (pp. 11- 39). Heidelberg, Germany: Springer. 1985

Ajzen, I., Madden, T. & Ellen, P. S. "A comparison of the Theory of Planned Behavior and the Theory of Reasoned Action," Personality and Social Psychology Bulletin, vol. 18(1), pp. 3-9. 1992

Anand, p. "Foundations of Rational Choice Under Risk", Oxford, Oxford University Press 1993 repr 1995 2002

Arthur, J.G. and Sonander, J. Safe to Use? Utlitity Tensions in Usability Engineering to Enhance Air Traffic Control Technology Safety. In proceedings HCI Aero 2000, Safety and Usability Concerns in Aeronautics IFIP WG workshop. 2000

Bakan, J., The Corporation: The pathological pursuit of profit and power. Constable, London. 2005

Baverstam, U., Decision Support and Decision Support Systems. Radiation Protection Dosimetry vol. 73, Nos. -4, pp. 1-6 (1997)

Beck, U (1986). Risikogesellschaft; Auf dem We in eine andere Moderne. Suhrkamp Verlag, Frankfurt am Main.

Beck, U (1992). Risk Society: Towards a New Modernity (translated by Mark Ritter). Sage, London.

Bell, D.E., Raiffa, H. and Tversky, A., (eds) Decision Making. Cambridge University Press. 1988

Boholm, A. Comparative Studies of risk perception: a review from twenty years of research. Journal of Risk Research 1 (2) 135-163. 1998

Carter S. Boundaries of danger and uncertainty: an analysis of the technological culture of risk assessment. In: Gabe J, editor. Medicine, health and risk: sociological approaches. Oxford: Blackwell; 1995. Chapter 7, p. 133-50. (WHO, 2002)

Chater, N. How smart can simple heuristics be? Open peer commentary on Todd, P.M. and Gigerenzer, G. Precis of simple heuristics that make us smart. Behavioural and Brain Sciences, 23, 745-746. 1999.

Chater, N., Oaksford, M., Ramin, N. and Redington, M. Fast, frugal, and rational: How rational norms explain behaviour. Organisational Behaviour and Human Decision Processes. 90 63-86. 2003.

Chater, N., Redington, M., Nakisa, R. & Oaksford, M. Rationality the fast and frugal

way. In M. G. Shafto and P. Langley (Eds.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, (pp. 96-101). Mawah, NJ: Erlbaum. 1997.

Cheng, P.W. and Holyoak, K.J. (1985) 'Pragmatic Reasoning Schemas', Cognitive Psychology 17: 391 – 416 (Oaksford and Chater 1993)

Cohen, L.J. Can human irrationality be experimentally determined? Behavioural and Brain Sciences, 4, 317 – 331. 1981 (O'Hagan et al, 2006)

Cooke, R.M. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press. (1991)

De Finetti, B., Theory of Probability: A Critical Introductory Treatment. Vol 1. J. Wiley and sons. 1974.

Dodd, L., Moffat, J. and Smith, J. Discontinuity in decision-making when objectives conflict: a military command decision case study. Journal of the Operational Research Society (2006) 57, 643-654

Douglas, M. Wildavsky, A. Risk and Culture: an essay on the selection of tehnological and environmental dangers. University of California Press. 1982

Dowie, J. Against risk. Risk, Decision and Policy 4 (1), 57-73. 1999

Dowie, J. Communication for better decisions; not about 'risk'. Health, Risk and Society, 1(1), 41-53. 1999

Doyle, J.K. (1997) Judging Cumulative Risk. Journal of Applied Psychology, 27, 6, 500-524

Englander, T., Farago, K., Slovic, P., Fischoff, B. (1986) A Comparative Analysis of Risk Perception in Hungary and the United States. Social Behaviour, 1, 55 – 66.

European Organisation for the Safety of Air Navigation. Investigating the Air Traffic Complexity: Potential impacts on workload and costs. EEC Note No. 11/00 project GEN-4-E2. 2000

Evans, J. St B.T. (1989) Bias in Human Reasoning; Causes and Consequences, London; Erlbaum (Oaksford and Chater, 1993)

Faucheaux, C. Cross-cultural research in experimental social psychology. European Journal of Social Psychology, 6, 269-322. 1976. (Cited in Boholm, 1998)

Fiedler, K. The dependence of the conjunction fallacy on subtle linguistic factors. Psychological Research, 50, 123 – 129. 1988 (in O'Hagan et al, 2006)

Fiedler, K. & Schmid, J. Heuristics. In A. S. R. Manstead, M. Hewstone, S. T. Fiske, M. A. Hogg, H. T. Reis and G. R. Semin (Eds.), The Blackwell Encyclopaedia of Social Psychology. Blackwell. 1995.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S. L., and Keeney, R. L. (1981). Acceptable Risk. New York: Cambridge University Press.

Fischoff, B. Risk Perception and Communication Unplugged. Twenty years of

process. Risk Analysis, 15, 2. 1995

Fischoff, B., Slovic, P. and Lichtenstien, S. Subjective Sensitivity Analysis. Organisational Behaviour and Human Decision Making Processes. June, 23(3), 339 – 359. 1980

Fischoff, B., Slovic, P. and Lichtenstien, S. The Public Versus 'The Experts'. In V.T. Covello, W.G. Flamm, J.V. Rodricks and R.G. Tardiff (Eds). The Analysis of Perceived Risks. New York, Plenum, 235 – 249. 1983

Fischoff, B., Slovic, P. and Lichtenstien, S. How Safe is Safe Enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sciences, 9, 127 – 152. 1978

Fishbein, M. & Ajzen, I. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research, Reading, Mass: Addison-Wesley. 1975.

Fisk, J.E. Conjunction Fallacy. In R.F. Pohl (Ed), Cognitive Illusions: A handbook on fallacies and biases in thinking, Judgement and Memory. Psychology Press. Hove. UK 2004 (O'Hagan et al, 2006)

Flemming, M., Flin, R., Mearns, K., and Gordon, R. Risk perception of offshore workers on UK oil and gas platforms. Risk Analysis, 18, 1. 1998

French, S. Group consensus probability distributions: a critical survey. In Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M. (eds) Bayesian Statistics 2, 183-201. 1985

Gifford S. The meaning of lumps: a case study of the ambiguities of risk. In: Stall R, Janes C, Gifford S, editors. Anthropology and epidemiology. Interdisciplinary approaches to the study of health and disease. Dordrecht: Reidel Publishing; 1986. p. 213-46. (WHO, 2002)

Gigerenzer, G. The Psychology of good judgment: Frequency formats and simple algorithms. Medical Decision Making, 16, 273 – 280. 1996 (in O'Hagan et al, 2006)

Griffin, D, Buehler, R. Frequency, Probability and Prediction: Easy solutions to cognitive illusions? Cognitive Psychology, 38, 48-78. 1999 (O'Hagan et al, 2006)

Golstein, M., and Wooff, D. Bayes Linear Statistics: Theory and Methods, Wiley. 2008

Goodwin, P., and Wright, G. Decision Analysis for Management Judgment, Wiley. 2008

Hogarth, R.M. Cognitive processes and the assessment of subjective probability distributions. JASA, 70, 271-294. 1975 (O'Hagan et al, 2006)

Hogarth, R.M., Judgment and Choice. J. Wiley and Sons, Chichester. 1987

HSE (1999). Risk perception and risk communication A review of literature. Health and Safety contract research report 248/1999

Inhelder, B. and Piaget, J. The Growth of Logical Reasoning, New York; Basic

Books. 1958 (Oaksford and Chater 1993)

Janz, N.K. and Becker, M.H. The Health Belief Model: A decade later. Health Education Quarterly, Spring, 11, 1 – 47. 1984

Johnson-Laird, P.N. and Byrne, R.M.J. Deduction, Hillsdale NJ; Erlbaum. 1991 (Oaksford and Chater 1993)

Jones, S.K., Jones, T.K., Frisch, D. Biases of probability assessment: a comparison of frequency and single case judgements. Organisational Behaviour and Human Decision Processes, 61, 109 – 122. 1995 (O'Hagan et al, 2006)

Kadane, J., Larkey, P. The confusion of IS and OUGHT", Management Science, 1983

Kadane, J., Larkey, P. "Subjective probability and the theory of games" Management Science, 1982

Kahneman, Daniel and Amos Tversky ([1979] 1988), "Prospect Theory: An analysis of decision under risk", pp. 183-214 in Gardenfors and Sahlin. 1988

Kahnemann, D., Slovic, P. and Tversky, A. (eds) Judgement Under Uncertainty. Cambridge University Press. 1982

Keeney, R.L. Value Focussed Thinking. Harvard University Press. London. 1992

Klein, G. Sources of Power: How people make decisions. MIT press, Cambridge, Mass. 1999

Koehler, J.J. The base rate fallacy re-considered: Descriptive, normative and methodological challenges. Behaviour and Brain Sciences, 19, 1 - 53. 1996 (O'Hagan et al, 2006)

Korthals, M. Ethics of Differences in Risk Perception and Views on Food Safety. Food Protection Trends 24(7), 498-503. 2004

Lichtenstein, S., Slovic, P., Fischoff, B., Layman, M. & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 551-78.

Lightowlers, P. Still Dirty: A review of action against toxic products in Europe. WWF-UK. 2004

Lindberg, G., Frost, D.E. (1992) Counterfactuals in Financial Decision Making. Acta Psychologica, 79, 227-244.

MacGregor, D.G. and Flemming, R. (1996). Risk Perception and Symptom Reporting. Risk Analysis, 16 (6)

McDermott, D. A Critique of Pure Reason, Technical Report, Department of Computer science, Yale University, June, 1986 (Oaksford and Chater 1993)

McDermott, D. A Critique of Pure Reason. Computational Intelligence, 3, 151-160. 1987

Meyer, A. M., Booker, J.M. Eliciting and Analyzing Expert Judgement. A Practical Guide. Academic Press Limited, London. 1991

National Air Traffic Services: Strategic Plan For Safety: 2001. NATS 2001

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J. R., Garthwhaite, P.H., Jenkinson, D.J., Oakley, J.E. and Rakow, T. Uncertain Judgements. Eliciting Experts' Probabilities. John Wiley and Sons. Chichester. 2006

Oaksford, M., & Chater, N. Commonsense reasoning, logic and human rationality. In R. Elio (Ed.), Commonsense reasoning and rationality. Oxford: Oxford University Press. 2002.

Oaksford, M., & Chater, N. The probabilistic approach to human reasoning. Trends in Cognitive Sciences. 5, 349-357. 2001.

Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K. I. Manktelow, & D. E. Over (Eds.), Rationality, (pp. 31-60). London: Routledge.

Oliver, R.M. and Smith, J.Q. (Eds.) Influence Diagrams, Belief Nets and Decision Analysis, Wiley. 1990

Perry, T.S. In Search of the Future of Air Traffic Control. IEEE Spectrum, August, p19-35. 1997

Peterson, C.R., Beach, L.R. man as intuitive statistician. Psychological Bulletin, 68, 29-46. 1967 (O'Hagan et al, 2006)

Phillips, L.D. A theory of requisite decision models. Acta Psychologica 56, 29-48. 1984

Phillips, L.D., Requisite Decision Modelling For Technological Projects. Vleck, Ch., Cvetkovitch, G. (eds) Social Decision Methodology for Technological Projects. 95 – 110. Klewer Academic. 1989

Pidgeon N. Risk perception. In: Royal Society. Risk analysis, perception and management. London: Royal Society; p. 89-134. 1992 (WHO, 2002)

Pidgeon, N. (1992) The Psychology of Risk. In Blockley, D. I. (ed) Engineering Safety. McGraw-Hill. Maidenhead.

Plous, S. *The psychology of judgment and decision making*. McGraw-Hill Inc. 1993.

Powel, D. and Leiss, W. Mad Cows and Mothers Milk: The perils of poor risk communication. McGill-Queens University Press, Montreal. 1997

Rappaport, R.A. Risk and the human environment. Annals of the American Academy of the Political and Social Sciences 545, 64-74. 1996 (Cited in Boholm, 1998)

Rasmussen, J. Human Errors; a taxonomy for describing human malfunction in industrial installations. Journal of Occupational Accidents, 4, 311-333, 1982

Ratzan, S. Making Sense of Risk (editorial). Journal of Health Communication 8(5), 399-400. 2003

Reason, J. Managing the Risks of Organisational Accidents. Ashgate, Aldershot. 1997

Rosenhead, J. (ed) Rational Analysis for a Problematic World. Wiley, Chichester. 1989

Royal Society, Study Group on Risk Assessment, Analysis, Perception and Management. Royal Society, London. 1992

Savage, L.J., The Foundations of Statistics. J Wiley & Sons. New York. 1954

Schwartz, N., Strack, F. Context effects in attitude surveys; Applying cognitive theory to social research. In W. Stroebe and M. Hewstone (Eds.), European Review of Social Psychology, 2, 1 – 50. 1991 (O'Hagan et al, 2006)

Sears, D.O. College sophomores in the laboratory: Influence of a narrow database on psychology's view of human nature. Journal of personality and social psychology, 51, 513 – 530. 1996 (O'Hagan et al, 2006)

Shafer, Glenn. Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning 3 1-40. 1990

Slovic P. The perception of risk. London: Earthscan; 2000. p. 473. 2000

Slovic, P. Perception of Risk. Science, vol 236, April, 280 – 285. 1987

Slovic, P., Fischoff, B. and Lichtnstein, S. Informing the public about the risks from ionising radiation in H.R. Arkes and K. R. Hammond (eds). Judgement and decision making - an interdisciplinary reader. Cambridge University Press. 280 – 285. 1986

Slovic, p., Finucane, M., Peters, E., MacGregor, D.G. : The Affect Heuristic In T. Gilovich D. Griffin, & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 397-42Press, 2002).

Smith, J.Q. Decision Analysis A Bayesian Approach. Chapman and Hall. 1988

Stern PC, Fineberg HV, editors National Research Council, Committee on Risk Characterisation.. Understanding risk. Informing decisions in a democratic society. Washington (DC): National Academy Press; 1996. (WHO, 2002)

Stirling, A. Risk, at a turning point? Journal of risk research, 1, 27, 97-109. 1998

Stirling, A. Deliberative mapping. In press. 2007

Strecher, V. J., and Rosenstock, I. M. "The Health Belief Model." In Health Behavior and Health Education: Theory, Research, and Practice, eds. K. Glanz, F. M. Lewis, and B. K. Rimer. San Francisco: Jossey-Bass. 1997

Tversky & Kahneman, 1974, p. 1124 cited in Plous, 1993

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1127 - 1131.

Tversky, A. & Kahneman, D. The Framing of Decisions and the Psychology of Choice. *Science, 211*, 453 - 458. 1981

Tversky, A. & Kahneman, D. Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press. 1982

Tversky, A.& Kahneman, D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207-232. 1973

Van der Plight, J. Perceived risk and vulnerability as predictors of precautionary behaviour, Brittish Journal of Health Psychology, 3, 1 – 14. 1998

Van der Velde, F.W., Van der Plight, J. and Hooijkaas, J. Perceiving AIDS related risks: accuracy as a function of differences in actual risk. Health Psychology, 13, 25 – 33. 1994

Von Mises, R. Probability Statistics and Truth. New York. 1939.

World Health Organisation (2002). World Health Report 2002

Wright, G. and Ayton, P. (eds) Subjective Probability. John Wiley and Sons. 1994

Wynne, B. Misunderstood Misunderstanding: Social Identities and Public Uptake of Science. Public Understanding of Science 1, 281-304. 1992