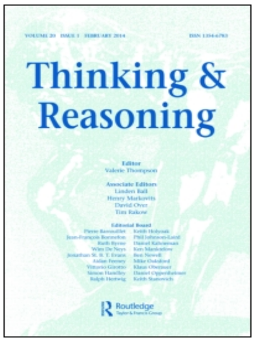

Routledge
Taylor & Francis Group

Thinking & Reasoning

ISSN: 1354-6783 (Print) 1464-0708 (Online) Journal homepage: https://www.tandfonline.com/loi/ptar20

Expert Judgement and Expert Disagreement

Jeryl L. Mumpower & Thomas R. Stewart

Expert Judgement and Expert Disagreement

Jeryl L. Mumpower and Thomas R. Stewart
University at Albany, USA

As Hammond has argued, traditional explanations for disagreement among experts (incompetence, venality, and ideology) are inadequate. The character and fallibilities of the human judgement process itself lead to persistent disagreements even among competent, honest, and disinterested experts. Social Judgement Theory provides powerful methods for analysing such judgementally based disagreements when the experts' judgement processes can be represented by additive models involving the same cues. However, the validity and usefulness of such representations depend on several conditions: (a) experts must agree on a problem definition, (b) experts must have access to the same information, and (c) experts must use the same organising principles. When these conditions are not met, methods for diagnosing and treating disagreement are poorly understood. As a start towards developing such an understanding, sources of expert disagreement are discussed and categorised.

INTRODUCTION

This paper examines expert judgement and expert disagreement, with special emphasis on public policy problems that contain a scientific or technical component. For many policy problems, the facts do not "speak for themselves". We do not know as much as we would like to know about the long-term health effects of carbon monoxide or the environmental effects of potential global warming, but public policy decisions on such issues cannot be deferred until we have perfect information; we never will. As definitive evidence is unavailable, we turn to scientific and technical experts to go beyond the data, to make inferences about the nature and severity of our problems, and to recommend potential solutions to them.

Why do experts in these circumstances so often disagree? This question is fundamental. If we understood the disagreement better, then we might be able to devise better ways to reduce it or cope with it. In addition to its practical significance, the issue has intrinsic intellectual interest. Because experts

Requests for reprints should be sent to Jeryl L. Mumpower, Center for Policy Research, University at Albany, State University of New York, Albany, NY 12222, USA.

Preparation of this paper was supported in part by the National Science Foundation (Division of Social and Economic Science) under Grant SES90-11868.

presumably exhibit the highest levels of judgemental ability, expertise has been and is a prominent topic in the behavioural study of judgement and decision making processes (Shanteau & Stewart, 1992).

The first step in improving our understanding of disputes among experts is to develop better means for describing and classifying different sources of disagreement. The present paper proposes such a framework. On the basis of research to date, certain portions of the framework can be specified in detail, whereas others can be only sketchily described.

After specifying a working definition of “expert judgement”, we identify three traditional explanations for disagreement. We then describe an alternative view that explains at least some instances of expert disagreement in terms of fundamental characteristics of human judgement processes. A well-understood and extensively studied framework for analysing disagreement among experts is then discussed. Finally, we propose a scheme for classifying reasons for disagreement among experts, and identify several major sources of disagreement for which we know little about either how to diagnose or treat the problem.

DEFINITION OF EXPERT JUDGEMENT

This paper focuses on disagreements among scientific and technical experts regarding inferences or predictions about the state of some environmental criterion variable that is, at least in principle, independently and empirically observable. For such judgements, expertise is ideally evaluated in terms of the degree of correspondence between the experts’ judgements and the environmental criterion.

In some fields of expertise, the environmental criterion may be directly observable, and objective evaluations of expertise are feasible. The performance of weather forecasters, for example, is routinely evaluated against observed weather during the forecast period (Murphy & Daan, 1985). Similarly, under certain conditions, medical diagnoses can be evaluated against “gold standards”, e.g. surgical observations or biopsy results (e.g. Getty, Pickett, D’Orsi, & Swets, 1988).

In many cases, however, the criterion is not readily observable. Sometimes, it is difficult to observe the environmental criterion because of time delays in feedback mechanisms. For example, if nothing is done, future developments will eventually prove whether draconian or conservative predictions about the effects of global warming are correct, but the answer will not come for decades. Sometimes, action taken because of expert forecasts makes it impossible to determine if the original predictions were correct. As the nations of the world take steps to implement the Montreal Protocol and reduce the level of emissions of chlorofluorocarbons (CFCs), the original predictions about effects on stratospheric ozone, based on the assumption of unabated emissions of

CFCs, become unverifiable. Still other times, there are ethical constraints that preclude the verification of expert judgements. It would be immoral to expose persons to levels of carbon monoxide suspected to cause long-term health effects in order to scientifically verify hypothesised mechanisms.

The definition of expertise that we use in this paper excludes circumstances in which an environmental criterion cannot be specified, even in theory. Claims of expertise in fields such as philosophy, art, ethics, literature, or mathematics, must necessarily be based on measures other than the correspondence of the judgement with the environmental criterion (e.g. consensus, coherence, or command of a factual knowledge base).

This paper emphasises the tradition of research on judgement and decision making that has been conducted primarily by psychologists and economists (Arkes & Hammond, 1986), and which has focused on describing human judgement processes in situations involving uncertainty, complexity, and trade-offs among multiple criteria. We will not attempt to consider the substantial bodies of research on expertise that have developed within other important intellectual traditions. These include traditional cognitive science, which has emphasised expertise in problem solving (e.g. Chi, Glaser, & Farr, 1988; Hoffman, 1992; Newell & Simon, 1972); research on the development, education, and socialisation of experts (e.g. Constant, 1989; Ladd & Lipset, 1972; Terman, 1954); and the extensive body of research on the politics and sociology of science (e.g. Cozzens & Gieryn, 1990; Jasanoff, 1987, 1990, 1993).

Finally, a key issue in discussions of expertise is deciding who qualifies as an expert. Although distinctions are made in many fields between experts, journeymen, novices, and beginners, there is no single valid indicator of scientific expertise relevant to policy problems that is comparable to, say, the ranking of chess masters or tennis players. To establish a workable starting place for examining expert disagreements, we follow the advice of Shanteau (1992) and will define experts to be those who are regarded as such by others within their field.

TRADITIONAL EXPLANATIONS FOR EXPERT DISAGREEMENT

Hammond (1996) has identified three traditional explanations for expert disputes and disagreements—incompetence, venality, and ideology. The *incompetence* explanation calls into question the qualifications, credentials, or intelligence of purported experts. According to this view, disagreements result because at least one disputant is not really considered an expert. For example, many physicians discount chiropractors' advice on medical matters because of what they regard as a lack of requisite training. A variation on this theme is that an expert, although perhaps fundamentally competent, did not invest the effort required to reach an informed judgement because of lack of time, resources, or motivation. Marcus (1988), for example, argued that lack of resources led to poor performance by toxicologists in the pesticide division of the Environmental Protection Agency, who were overtaxed by the need to re-register more than 50,000 substances.

The *venality* explanation views disagreements among experts in the same way as other garden-variety disputes—grounded in differences in personal interests. According to this view, experts take positions that serve their immediate self interests, or the interests of others who have the power to reward them. Experts working for businesses, for example, are accused of reaching conclusions that favour their employers' financial interests. Bell (1992) has documented a number of cases where venality appeared to play a role in scientific controversies. He claims that cases of political influence, conflict of interest, and fraud are indicators of general corruption in the scientific establishment. Others, such as Koshland (1987), have argued that highly publicised cases of misconduct reflect the actions of a tiny minority of scientists.

The *ideology* explanation assumes that experts' positions on policy-relevant scientific or technical issues are determined by political, religious, or ethical values or beliefs. According to this explanation, experts form opinions based on their personal sympathies, then select evidence that supports their opinions. For instance, atmospheric scientists with environmentalist sympathies might seek and find evidence to suggest that global warming is an imminently serious problem in order to justify environmental objectives, such as energy conservation, that they support on other grounds. A similar explanation has been applied to differences in opinion about nuclear power (Mitchell, 1984, p.l43).

JUDGEMENTAL BASES FOR EXPERT DISAGREEMENT

Expert disagreements may be attributable to the character and fallibilities of human judgement itself, rather than to incompetence, self-interest, or personal values. In short, experts may disagree because they *think* about the problem differently. When confronted with multidisciplinary issues involving scientific complexity and uncertainty, even competent, honest, and disinterested scientists may arrive at different conclusions.

This cognitive explanation for disagreement among experts originated with the work of Hammond and his colleagues, who showed that Brunswik's lens model, Social Judgement Theory (SJT), and Cognitive Continuum Theory (CCT) provide tools for understanding interpersonal disagreements (Brehmer, 1976a; Hammond, 1973) and helping to resolve conflict. Hammond and Adelman (1976) extended the interpersonal conflict paradigm into the domain of expert judgement and argued that SJT could assist in public policy disputes involving disagreements among experts. Hammond, Anderson, Sutherland, and

Marvin (1984) provided a further demonstration of the value of SJT and CCT in helping scientists improve their judgements of risk, and Hammond, Hamm, Grassia, and Pearson (1987) demonstrated that CCT could be useful in analysing the judgement processes of expert highway engineers engaged in different types of judgement tasks. Additional studies of disagreement among experts within the SJT framework have been done by Adelman and Mumpower (1979); Chaput de Saintonge and Hattersley (1985); Mumpower, Livingston, and Lee (1987); Stewart et al. (1989); Stewart, Moninger, Heideman, and Reagan-Cirincione (1992); Kirwan, Barnes, Davies, and Currey (1988); Wigton (1988); and Lusk, Stewart, Hammond, and Potts (1990); among others.

In the following section, we discuss a major source of judgement-based disagreement that is well understood, and briefly describe the techniques that have been used to deal with it. Later, we describe some sources of disagreement that are less well understood.

A "WELL UNDERSTOOD" CASE: DISAGREEMENT AMONG EXPERTS THAT CAN BE MODELLED BY ADDITIVE MODELS

Judgement-based disagreement between experts may result from either systematic or non-systematic differences in judgement processes. Systematic ones are relatively stable differences in how people integrate information into a judgement. Non-systematic differences refer to inconsistency or unreliability that introduces a random component into the process.

Systematic Differences in Judgement Processes

In SJT, Brunswik's (1952) lens model provides the theoretical framework for the analysis of expert judgement (see Cooksey in this issue). The application of this framework requires that experts (a) judge the same environmental criteria, and (b) have access to the same cues. In this context, two experts, *j* and *k*, may reach different judgements ($Y_{sj} \neq Y_{sk}$) because of systematic differences in the way that they integrate the information provided by the cues (X_i). Such differences will be reflected in one or more of four ways:

1. *Different organising principles*. Two experts may disagree concerning the appropriate way to combine information.
2. *Different weights*. Experts may believe that different pieces of information are more important.
3. *Different function forms*. Two experts may disagree concerning the appropriate functional relationship between levels of cues and judgements.
4. *Differences in bias*. Even if two experts use the same organising principle and their relative weights and function forms are identical, they may still make different judgements because of differences in the mean and variance of their judgements.

Research to date in the SJT framework has concerned primarily differences in weights and function forms. Because it has generally been assumed that the organising principle is additive (see Hammond et al., 1984, for an exception), differences in organising principle have not been widely studied. As correlations (which are not affected by differences in the means and standard deviation of judgements) are typically used to measure agreement, the contribution of bias to disagreement has not been a major focus of research, although some studies have examined differences in means and standard deviations of judgements (e.g. Ullman & Doherty, 1984). Stewart (1990) has shown how bias can be incorporated into lens model analyses of expert disagreement.

Reasons for Systematic Differences in Judgement Processes

How can competent, honest, and disinterested experts differ in their judgement policies? Several factors may account for the development and persistence of differences in expert judgement. These factors may be grouped into four classes: (a) poor quality or missing feedback, (b) poor quality or missing information, (c) inability to learn about the quality of one's own judgement, and (d) causal texture of the environment.

Missing or Poor Feedback. A substantial body of research indicates that it is difficult to learn from experience when that experience comes in the form of outcome feedback concerning whether or not one's judgement was correct (Brehmer, 1980; Klayman, 1988). This is especially true when such feedback contains noise or error; learning under conditions of uncertainty has been shown to be extraordinarily difficult (Brehmer, 1976b).

The real world often provides amorphous feedback, however, so that witnesses do not agree about its meaning or interpretation. For example, some models show that the rise in global temperature during the last century should have been twice the 0.5ºC. observed. To sceptics, this gap casts doubt on the credibility of the models. To others, the rise is confirmatory evidence for the validity of the model (Stevens, 1989). For scientists, as well as laypersons, there is a tendency to interpret ambiguous or uncertain data so that it supports pre-existing beliefs (Mitroff, 1983; National Research Council, 1989, p.38ff). New confirmatory data are more readily accepted; contradictory information is more readily dismissed.

Missing or Poor Data. Often, judges do not have access to the data that they would like, but must rely on proxy variables to stand in for more relevant, but unavailable data. Research suggests that it is cognitively difficult to assign appropriate weights to proxy variables (Fischer, Damodaran, Laskey, & Lincoln, 1987). Being forced to rely on proxy variables that are only crudely related to the variable of true interest might exaggerate differences between experts and encourage reliance on irrelevant factors, as observed by Gaeth and Shanteau (1984).

Difficulty in Evaluating the Quality of One's Own Judgement. Einhorn and Hogarth (1978) argued that, in general, it may be difficult for people to make accurate appraisals of their own judgemental abilities. Under many common circumstances, people tend to overestimate the quality of their judgement because: (1) they tend to evaluate their expertise on the basis of the number of times they recall correctly predicting a positive outcome, rather than on a statistically appropriate basis that takes into account negative outcomes and failures to predict positive ones; (2) if there is a high base rate of true positives, then the ratio of true positives to false positives will be high, even if the judge has no discriminatory judgemental ability; (3) whenever the selection rate is higher than the base rate, little judgemental ability is required to obtain high frequencies of true positives; and (4) self-fulfilling treatment effects may inflate the observed number of true positives, leading to substantial overestimation of judgemental abilities. If expert judges believe they are better than they really are, they have little incentive to change or to try to reconcile their judgements with those of others.

Causal Texture of the Environment. Intercorrelated cues can hide disagreement because they may lead different experts to the same judgement even though they have distinctly different judgement policies (Mumpower & Hammond, 1974). This may occur, for example, when one person relies exclusively on cue A, while the other person relies exclusively on cue B. If A and B are highly correlated, then either cue will usually lead to the same judgement. Because of the resulting "false agreement", such differences may persist unrecognised. If cases arise in which the relevant cues do not covary in the usual pattern, however, differences in judgement policies may suddenly become apparent when the experts unexpectedly reach quite different judgements.

Non-systematic Differences in Judgement Processes

Random error is a second possible source of expert disagreement. Hammond and Summers (1972) introduced the concept of *cognitive control* to account for the fact that judges often do not integrate multiple cues in a perfectly predictable fashion. Brehmer and Brehmer (1988) observed that decades of research on modelling judgement processes have routinely shown that it is impossible to account for all the variance in individuals' judgements solely in terms of the cues, at least when the judgements have an intuitive component. Part of this lack of predictability is caused by judgemental unreliability. The exercise of judgement contains a skill component associated with the ability to apply knowledge about cue–judgement relationships (Brehmer, 1971). Research in applied settings has found that judgements by experts from a variety of disciplines frequently demonstrate a lack of perfect predictability, as well as substantial degrees of inaccuracy.

Research indicates that judges are most likely to exhibit imperfect cognitive control when they face environmental systems that are imperfectly predictable, when accurate judgements require use of large numbers of cues, when the relationship between cues and judgements is nonlinear, and when the judge is learning to cope with a new or novel judgement problem or learning a new way to make judgements about a familiar task (Brehmer 1976b, 1978; Einhorn, 1971; Harvey, 1995; Faust, 1986; Hogarth, 1987; Stewart et al., 1992). The conditions that contribute to judgemental unreliability seem to characterise most problems that combine scientific, technical, and policy components.

Random error in the judgement process is not a potential source of error, of course, when the judgement process is based on explicit, step-by-step rules, consistently applied. Analytical modes of thinking are generally less susceptible to random error, although they are not completely immune. Examples of failures to execute even simple analytical processes perfectly consistently (e.g. the non-balanced chequebook) are easy to cite.

ILL-UNDERSTOOD REASONS FOR JUDGEMENTAL DISAGREEMENT

Differences in weights and function forms have been the most extensively studied source of disagreement between experts. Meaningful analysis of disagreement in such terms, however, requires substantial prior agreement among experts about certain fundamental issues, including the nature of the problem and its structure, relevant information, and the basic type of organising principle appropriate for addressing the issue. In this section, we discuss the implications of disagreements among experts about these fundamental issues.

First, however, we need to introduce the idea of a judgement structure, which often takes the form of a hierarchy describing how concrete data are organised into intermediate judgements which are, in turn, organised into higher-level judgements.

Judgement Structure and Expert Disagreement

Expert judgement typically involves many cues that are organised into judgements in several steps. To form an opinion, experts must often select the most important and relevant theory and results from a variety of sources. Expert judgements about the value or state of environmental criteria almost always depend on prior judgements. For instance, to reach a judgement about the greenhouse effect, scientists may rely on information from the climatological record, empirical research, and atmospheric modelling. They then must weigh that information and aggregate it into judgements that are the basis for their conclusions.

The information on which such a judgement is based can often be organised into a structure that takes the form of a multi-level hierarchy (Stewart, West, Hammond, & Kreith, 1975). Global, policy-relevant judgements are at the top level. The lowest level of the hierarchy includes the most "objective" information available. Intermediate levels consist of judgements based on information at lower levels of the hierarchy. We propose that, for many problems, differences in the experts' judgements are attributable to differences in judgement hierarchies.

Stewart (1991) has argued that in the case of scientific judgement regarding global warming, information and intermediate judgements can be organised into a judgement hierarchy with at least five levels. Level 1 includes specific studies and results, as well as well-established theory. This level consists of the raw data or "facts" on which judgements are based. The elements of level 1 are "objective" in the sense that they are not controversial. In other words, level 1 includes the data about which scientists agree. It is the level of consensus.

Level 2 consists of a grouping of studies and results. As the studies at level 1 may number in the thousands, it is necessary to reduce and simplify them by grouping related studies in some way. This may be done through influential review papers or books, or by combining the work of a particular investigator or laboratory (e.g. "Smith's work shows that . . ."). There may be some disagreement among scientists regarding the most useful method for grouping level 1 data into level 2 and about where the weight of evidence within a particular grouping falls.

Level 3 includes the interpretation and aggregation of particular lines of research, e.g. detection of global warming in the climate record, measurement and prediction of trends in emissions of greenhouse gases, or results of global climate models. Significant disagreement is expected at this level.

Level 4 is for broad conclusions based on the evidence included in the lower levels. This level might include conclusions about the rate of global warming, the distribution of climatic changes, and the social and economic impacts of those changes. According to our definition, level 4 is the highest level of expert judgement.

Level 5 consists of policy recommendations based on the broad conclusions at lower levels. Disagreements at this level are not resolvable by scientific study alone because they involve questions of "what should be" that require value-based judgements about trade-offs between the costs and benefits of various actions and possible outcomes.

Stewart (1991) hypothesised that (1) scientific disagreement and uncertainty will be greater at higher levels of the hierarchy than at lower levels; (2) there will be individual differences among scientists with regard to their uncertainty about and confidence in their judgements both within and across levels; (3) there will be individual differences among scientists about how to select and aggregate information at lower levels to form a judgement at the next-higher level; and (4) these individual differences help to explain important disagreements at higher levels of the hierarchy.

Developing a structure for each expert can contribute to understanding the reasons for disagreements among experts. Indeed, analysis of expert disagreement should begin by developing such a structure for each expert. This approach is not unique to SJT. It has been used in decision theoretic and MAUT-based approaches for some time (Atman, Bostrom, Fischoff, & Morgan, 1994; Bostrom, Atman, Fischoff, & Morgan, 1994; Matzkevich & Abramson, 1995; von Winterfeldt & Edwards, 1986). Comparing the structures of different experts provides insights into reasons for differences in their judgements. The reasons for disagreement that are based in structure, rather than in differences in model parameters, we will call "structural disagreement".

Sources of structural disagreement can be classified into three broad categories: problem definition differences, information differences, and organising principle differences. These are summarised in Table 1.

TABLE 1
Categories and Sources of Structural Disagreement Among Experts, with Associated Diagnostic Questions

DIFFERENT PROBLEM DEFINITIONS
Fact–Value Confusion: *Does the disagreement concern solely scientific issues?*
Different Environmental Criteria: *Are experts judging the same environmental criterion?*
DIFFERENT INFORMATION
Differential Availability and Use of Information: *Do experts have available and make use of the same information?*
DIFFERENT ORGANISING PRINCIPLES
Qualitatively Different Mental Models: *Do experts have the same mental models of the environmental process? Do they agree about the identity of variables, the direction of causality, and the primacy of causal factors?*
Different Modes of Cognition: *Are the experts located at the same place along the cognitive continuum (from intuitive to analytical processes)?*
Different Organising and Integrative Models: *Can the experts be represented by the same type of model (i.e. linear additive model, rule-based model, non-linear algebraic model, dynamic feedback-based model, etc.) ?*

Disagreement Due to Problem Definition

Experts may disagree because they define the judgement problem differently, and are therefore not making judgements about the same thing. The two major reasons for different problem definitions are fact–value confusion and use of a different environmental criterion.

Fact–Value Confusion. Disagreements among experts may occur because questions of fact are entangled with value issues. Disagreement may not reflect different scientific or technical judgements about "what is" or "what will be", but rather different conclusions about "what ought to be" or "what should be done". To detect fact–value confusion, it may be helpful to consider whether the disagreement involves a clear, agreed-on environmental criterion. If the criterion is missing or subjective (e.g. "the public interest", "well-being of future generations"), then the judgements are likely to have a strong value component. Even when there is an agreed-on criterion, nominally technical judgements may be confounded with values. For instance, many experts endorse a programme of active governmental interventions designed to forestall global warming, whereas some express serious scepticism about the wisdom of such a programme (Stewart, Mumpower, & Reagan-Cirincione, 1992). Such differences may be attributable to disagreements about the level of global warming that would result if no interventions were taken (a factual issue), to disagreements about appropriate societal behaviour in the face of such a risky prospect (a value issue), or both.

Fact–value confusion may not be consciously, or even unconsciously, motivated by a desire to influence policy results. Such confusion can take place even when experts sincerely aspire to objectivity, but are unable to achieve it because they unwittingly allow values to influence their judgements when the issue is poorly formulated or the evidence is unclear. A potential cause of such confusion is the desire that experts are likely to feel to be responsive to policy makers. When experts are specifically asked to address questions of "what ought to be" they are likely to respond, even though it is impossible for them to respond "objectively" or "scientifically". When questions of facts and values are intermixed, pressure mounts for policy makers to attempt to become "amateur scientists" and/or for scientists to assume the role of "*de facto* policy makers" (Hammond & Adelman, 1976; Hammond & Mumpower, 1979).

Much has been written about the role of facts and values in judgement and in the policy process. At one extreme is the view that facts and values cannot and should not be distinguished or separated (for an example of such a stance, see Jasanoff, 1990). At the other is the view that they must be separated, and responsibility for factual and value judgements rests with different actors in the policy-making process (Hammond & Adelman, 1976). Proposed remedies include science courts (Casper, 1976; Kantrowitz, 1967), mediation (Abrams & Berry, 1977), blue ribbon commissions, and citizen panels.

This debate is further obfuscated by the confusion surrounding the multiple senses in which the term “value judgement” is used. In addition to differences in pecuniary values (i.e. those linked to self-interest) and social values (i.e. those that reflect ethical, religious, or social beliefs), scientists may differ, sometimes vehemently, about the best ways to conduct the scientific enterprise. Disputes about methodology, the scientific process, and the nature of data and information are endemic to science (Hammond, 1987). Hammond (1996) has identified four persistent sources of methodological disagreement among scientists: (1) disagreement about whether irreducible or reducible uncertainty prevails, (2) subjectivist vs. frequentist views of probability, (3) whether research should seek confirmation or disconfirmation of hypotheses, and (4) correspondence vs. coherence theories of truth. Because scientists disagree about methodology, it is often said that science is always influenced by value judgements. It might be clearer to say that science is always influenced by *subjective* judgements. Subjective methodological judgements may indeed influence the conclusions that researchers reach, but not in order to serve narrow self-interests or to promote favoured social values. It is unfortunate that we often use the same term—value judgements—to describe both situations in which answers are influenced by ulterior, concealed motives and those in which answers are affected by subjective methodological or epistemological differences, which disputants typically go to great pains to try to make clear and explicit.

Our point of view—that to create a better basis for analysing and understanding expert disagreements, an attempt should be made to separate questions of scientific fact from those of personal and social value—is hardly unique (for an example of an explicit endorsement of this principle, see National Research Council, 1983). Although it may not be possible ordinarily to separate cleanly questions of fact from questions of value, the perfect should not be the enemy of the good. Difficult as it may prove in practice, we believe it is preferable to strive to maintain a distinction between scientific and value issues, rather than to intentionally and deliberately intermix the two types of questions.

Different Environmental Criteria. Even if the concerns are exclusively scientific, experts may disagree about the nature of the problem and the required judgements. If experts use identical terms to mean different things, they will make different types of judgements, perhaps without realising it. If they were to assess precisely the same phenomenon, they might agree, but, because they never make the same judgements, it is impossible to know for sure whether they really disagree.

For example, in 1991, as a result of reports of five people who became infected with HIV after treatment by a Florida dentist, there was public concern about healthcare professionals transmitting HIV to their patients and calls for AIDS testing of healthcare professionals. The controversy focused on the problem of transmission of HIV in healthcare settings. Some experts, noting that there was little evidence that patients have ever been infected through contact with infected healthcare workers, identified a potentially more serious problem: transmission of HIV from one patient to another through shared use of equipment and improper sterilisation (Rogers & Gellin, 1991). If two groups of experts had adopted these two different definitions of the problem (predicting risk of physician–patient transmission vs. predicting risk of patient–patient transmission), they would have sought different data to analyse the risk, arrived at different conclusions, and suggested different remedies. As their judgements would be validated by different environmental criteria, they could disagree markedly and still both be right. There would be no point in analysing the reasons for their disagreement until they agreed to act on identical formulations of the problem.

The reasons for disagreements about the appropriate definition of the environmental criterion may range from the profound to the trivial. Different problem definitions may be tied to differences in personal or social values, which indirectly lead the experts to focus on disparate aspects of the problem. Differences in problem definition may result from differences in disciplinary or scientific training or orientation that lead to strongly held beliefs about how science ought to done, including exactly what questions ought to be asked and how they should be answered. For instance, Schneider (1989) recounted the controversy that followed James Hansen's 1988 statement before the US Senate Committee on Energy and Natural Resources that "we can state with about 99% confidence that current temperatures represent a real warming trend rather than a chance fluctuation over the 30-year period." Although Schneider reported that he was in basic sympathy with Hansen's argument, he found himself repeatedly in the uncomfortable position of being asked to endorse or renounce Hansen's position, when it reflected a problem definition with which he was less than comfortable. He quoted his response to one such attempt: "I'm not going to get into a false dichotomy debate with Jim Hansen . . . True, I choose to state the problem differently than he has, but focusing only on that difference is not going to give the right impression."

Disagreement Due to Different Information or Data

Even if experts make judgements about precisely the same environmental criterion, they may reach different conclusions if they have access to, or make use of, different information or data. Sometimes identical data are not available to different experts for mundane reasons. Because new data are not necessarily disseminated instantaneously, some experts may not have access to the most recent relevant information, although this problem can normally be readily remedied, if so desired. Sometimes, however, there may be substantial incentives to intentionally withhold information. For instance, a pharmaceutical company involved in developing a new drug may not wish to share with competitors or regulators all relevant data concerning the drug's potential risks or benefits.

Even when two experts take similar approaches to the problem and have access to the same data set, the inputs to their judgement process may not be precisely identical. The distinction between data and judgement is often quite blurred. In many fields, there is no universally agreed-on corpus of "hard data" that can be input into competing models. Elements of judgement and interpretation are always involved in screening, selecting, interpreting, and aggregating the data that serve as inputs to expert judgement processes.

A further potential difficulty concerns deciding which data are relevant to the problem. Answers to these sorts of questions are often influenced by deeply held beliefs about methodology, the scientific process, and the nature of data and information. For instance, the field of academic psychology was greatly influenced for many years by arguments by Watson and other behaviourists that the only data relevant to the study of people were observable behaviours. This caused many psychologists to reject altogether the validity of introspective, participant-observer, and most cognitive data. Debates about the appropriate data for psychological science still persist, and similar disputes can be found in many other fields.

Such methodological differences were evident in a recent exchange in *Science*. Pool (1992, p.44) suggested that computer simulations were becoming so accurate that they "can take their place alongside experimental data as an object of scientific study". In a subsequent letter, Ross (1992, p.86) objected strenuously to the idea that discoveries can be made through the study of computer simulation. He argued that anything produced by a computer simulation is a *prediction* and "[t]he validation of a prediction is confirmation by experiment". The argument about the value of computer simulations vs. empirical experiments reflects deep-seated beliefs about how science should be conducted (see the discussion of correspondence vs. coherence theories of truth in Hammond, 1996).

Disagreement Due to Different Organising Principles

Even if experts make judgements about the same environmental criterion and have access to the same body of relevant information, they may organise the information in qualitatively different ways. Organising principles can differ at three levels. First, experts may have structurally different "mental models" of the pertinent process. They may disagree about the identity of variables, the direction of causality, or the primacy of causal factors. Second, experts may have the same structural models but use different modes of cognition to reach judgements. Some experts may approach the problem in a highly analytical fashion, whereas others are more intuitive. Third, experts may use the same modes of cognition, but employ different principles for organising information. For example, one expert might use an additive organising principle, whereas another uses a multiplicative one.

Different Mental Models. A key step in understanding instances of expert disagreement is to determine if the experts have qualitatively different mental models or ways of organising information. If this is not obvious from computer or other formal models developed by the experts themselves, eliciting structural models, influence diagrams, or explicit descriptions of the judgement hierarchy may help to clarify differences between experts' mental models of the process.

Judgements about complex environmental criteria are almost always enmeshed in a complex web of prior judgements, a reality that may frustrate attempts to get at the root of expert disagreements. The search for an explanation of disagreements will frequently require unfolding prior judgements, a process that may sometimes lead backwards in a seemingly infinite regress.

Different Modes of Cognition. Even if experts are similar in all other respects, they may use different types of cognitive processes to aggregate information into a judgement. In his Cognitive Continuum Theory (see the paper by Doherty and Kurz in this issue), Hammond (1980, 1981) drew a distinction between *intuitive* and *analytical* judgement. He proposed that judgement processes can be ordered along a continuum that is anchored at one pole by intuitive thinking and at the other pole by analytical, rational thinking.

Pure intuitive thinking tends to be holistic, to rely on weighing and combining multiple pieces of evidence, to be relatively fast, and to make use of comparatively small numbers of cues. Experts using intuitive judgement processes often lack self awareness of the details of the process and frequently rely on visual metaphors (Brunswik, 1956; Hammond, 1980, 1981). Pure analytical thinking relies on a step-by-step, rule-bound approach. It tends to be comparatively slow, to make use of large numbers of cues, and to involve greater conscious awareness as the problem is worked through explicit stages that typically provide for a clear retracing of process. It is more likely to rely on quantitative metaphors and tools such as paper and pencil, calculators, and computers.

Kerr's (1989) report on the debate between Lindzen and his various critics illustrates the potential tension between intuitive and analytical modes of thinking in a practical, policy context. Lindzen, who represents the intuitive approach to the global warming problem, describes his own idea about how control of atmospheric temperature works as "an idea of a theological or philosophical nature". Climate modellers have criticised Lindzen because he "seems to claim he has a better climate model in his head than they have in their supercomputers." Schneider, a proponent of a more analytical computer model-based approach, asked, rhetorically, "Does he have a calculation, or is his brain better than our models?" (Kerr, 1989, p.1118).

Most real-world problems evoke a mode of response that is neither purely intuitive or analytical, but involves a mixture of analysis and intuition that can be described as *quasirationality* (Brunswik, 1952, 1956; Hammond, 1987). Quasirational thinking involves alternations between, or combinations of, intuition and analysis. Hammond (1987) argued that most tasks requiring expertise are quasirational. In his view, experts are not required for intuitive tasks in which a person's judgement admits of no appeal to reportable and defensible reasoning. Nor are experts needed for analytical work that involves routine calculations. Expertise is ordinarily required for those tasks that demand some combination of intuitive and analytical abilities.

Generally, it will be difficult to make meaningful comparisons between judgement processes that are not located at approximately the same position along the cognitive continuum. If one expert relies on an analytical approach to the problem, whereas a second approaches the problem intuitively, it may not be useful to say anything more than that the two experts aggregate information in qualitatively quite different ways. In order to resolve such disagreements, experts would have to agree to use a common mode of thought—analytical thinkers would need to think intuitively, or vice versa.

Different Organising and Integrative Models. Even if experts use similar cognitive processes when they address the problem, they may not organise and integrate the information in the same way. Different types of models—algebraic, Bayesian, rule-based, dynamic, or other—may be best suited for representing different experts' judgement processes.

Expert judgement processes are not ordinarily accessible to direct observation. Models of those processes must be constructed. Sometimes experts construct their own models, for example, when climatologists develop elaborate general circulation models that describe their judgement about how levels of carbon dioxide and other atmospheric gases are related to phenomena such as global warming. Sometimes such models are developed by psychologists or others (e.g. "knowledge engineers") who wish to describe the judgement processes of experts.

An extraordinary variety of models can be used to describe expert judgement. Some common examples are:

- *Linear additive models*. Such models assume that the judgement process can be represented by an expression specifying how to weigh and combine multiple pieces of information. They are ordinarily most appropriate for describing processes towards the intuitive end of the cognitive continuum.
- *Non-linear additive models*. The relation between each cue and the judgement is a non-linear function of the cue (e.g. exponential, quadratic, etc.).

- *Non-linear algebraic models*. The cues are combined by some process other than addition (e.g. multiplication). Such models are generally most appropriate for quasirational judgement processes that are somewhat more intuitive in character.
- *If–then, rule-based models*. The judgement process is represented by a sequence of rule-bound steps. They are ordinarily most appropriate for describing processes towards the analytical end of the judgement-process continuum.
- *Dynamic simulation models*. These models are concerned with change over time, and typically contain feedback loops in which certain variables are modelled in reciprocal causal relationships with one another. These models tend to reflect a more analytical orientation towards the problem; they typically rely on difference equations, or similar quantitative techniques, for expressing relationships between variables.

Although there tends to be a correspondence between judgement processes and the types of models constructed of them, it is possible to represent highly analytical processes by simple linear models, or to approximate highly intuitive processes by an if–then, rule-based expert system, sometimes even achieving a substantial degree of predictive accuracy. This is possible because all judgement models are *paramorphic* (Hoffman, 1960), describing and representing (but not attempting to reproduce) some, but never all, of the relevant aspects of judgement processes. Evaluations of the appropriateness of different types of models for representing specific judgement processes may consider a variety of criteria in addition to predictive accuracy, including fidelity to process, parsimony, comprehensibility, and so on. The relative importance of such criteria will always be contingent on the circumstances and the purpose of the modelling process.

CONCLUSION

We have made three broad claims. First, following Hammond, we claim that judgemental processes are an important source of expert disagreement and that disagreement can be understood and, in many cases, resolved without recourse to explanations based on venality, ideology, or incompetence. Second, we claim that all reasons for judgementally based expert disagreement can be classified into several levels. Third, we claim that agreement on problem structure and organising principle is necessary before an analysis based on weights and function forms can come into play.

A now substantial body of research has demonstrated that SJT often offers a valuable means for analysing and understanding the sources of expert disagreement and that the results of SJT analyses may help to resolve disputes among experts. These successes do not imply, however, that *all* disagreements can be analysed in terms of differences in the weights and function forms associated with the terms in a linear model. In order for SJT to be appropriate and meaningful, extensive preliminary work must be done and a number of preconditions met. A common problem definition must be established; the same information must be available to all experts; and it must be ascertained that the experts share a similar problem structure, think about the problem at roughly the same point along the cognitive continuum, and employ the same type of organising principle. We do not yet know much about how to describe or analyse—much less, to help to resolve—differences between experts when these fundamental preconditions are not satisfied. This provides an important and exciting challenge for the future. By making clearer the various forms that disagreements among experts can and do take, we hope that the present analysis represents an initial step towards extending and further improving our understanding of expert disagreements.

REFERENCES

Abrams, N.E., & Berry, R.S. (1977). Mediation: A better alternative to science courts. *Bulletin of the Atomic Scientists*, *33*, 50–53.

Adelman, L., & Mumpower, J.L. (1979). The analysis of expert judgement. *Technological Forecasting and Social Change*, *15*, 191–204.

Arkes, H.R., & Hammond, K.R. (Eds.). (1986). *Judgement and decision making: An interdisciplinary reader*. Cambridge: Cambridge University Press.

Atman, C.J., Bostrom, A., Fischhoff, B., & Morgan, M.G. (1994). Evaluating risk communications: Completing and correcting mental models of hazardous processes: I. *Risk Analysis*, *14*, 779–788.

Bell, R. (1992). *Impure science: Fraud compromise and political influence in scientific research*. New York: Wiley.

Bostrom, A., Atman, C.J., Fischhoff, B., & Morgan, M.G. (1994). Evaluating risk communications: Completing and correcting mental models of hazardous processes: II. *Risk Analysis*, *14*, 789–798.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C.R.B. Joyce (Eds.), *Human judgment: The social judgment theory view*. Amsterdam: North-Holland.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*(6), 259–260.

Brehmer, B. (1976a). Social judgment theory and the analysis of interpersonal conflict. *Psychological Bulletin*, *83*(6), 985–1003.

Brehmer, B. (1976b). Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psychology*, *17*, 300–312.

Brehmer, B. (1978). Response consistency in probabilistic inference tasks. *Organizational Behavior and Human Performance*, *22*, 103–115.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, *45*, 223–241.

Brunswik E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd Edn.). Berkeley: University of California Press.

Casper, B.M. (1976). Technology policy and democracy: Is the proposed science court what we need? *Science*, *194*, 29–35.

Chaput de Saintonge, D.M., & Hattersley, L.A. (1985). *Who needs antibiotics for otitis media? Can we help doctors agree?* Unpublished manuscript, London Hospital Medical College, Department of Clinical Pharmacology and Therapeutics.

Chi, M.T.H., Glaser, R., & Farr, M.J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Constant, E.W. II. (1989). Science in society: Petroleum engineers and the oil fraternity in Texas, 1925–65. *Social Studies of Science*, *19*, 439–472.

Cooksey, R.W. (this issue). The methodology of social judgement theory. *Thinking and Reasoning*, *2*.

Cozzens, S.E., & Gieryn, T.F. (Eds.). (1990). *Theories of science and society*. Bloomington, IN: Indiana University Press.

Doherty, M.E., & Kurz, E.M. (this issue). Social judgement theory. *Thinking and Reasoning*, *2*.

Einhorn, H. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, *6*, 1–27.

Einhorn, H., & Hogarth, R.M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*, 395–416.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology*, *17*(5), 420–430.

Fischer, G.W., Damodaran, N., Laskey, K.B., & Lincoln, D. (1987). Preferences for proxy attributes. *Management Science*, *33*(2), 198–214.

Gaeth, G.J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, *33*(2), 263–282.

Getty, D.J., Pickett, R.M., D'Orsi, C.J., & Swets, J.A. (1988). Enhanced interpretation of diagnostic images. *Investigative Radiology*, *23*, 240–252.

Hammond, K.R. (1973). The cognitive conflict paradigm. In L. Rappoport & D.A. Summers (Eds.), *Human judgment and social interaction* (pp.188–205), New York: Holt, Rinehart & Winston.

Hammond, K.R. (1980). *The integration of research in judgment and decision theory* (Tech. Rep. No. 226): Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Hammond, K.R. (1981). *Principles of organization in intuitive and analytical cognition* (Report No. 231). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Hammond, K.R. (1987). *Reducing disputes among experts* (Tech. Rep. No. 268). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Hammond, K.R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hammond, K.R., & Adelman, L. (1976). Science, values, and human judgment. *Science*, 1*94,* 389–396.

Hammond, K.R., Anderson, B.F., Sutherland, J., & Marvin, B. (1984). Improving scientists' judgments of risk. *Risk Analysis*, *4*(1), 69–78.

Hammond, K.R., Hamm, R.M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*, 753–770.

Hammond, K.R., & Mumpower, J.L. (1979). Formation of social policy: Risks and safeguards. *Knowledge: Creation, Diffusion, Utilization*, *1*, 245–258.

Hammond, K.R. & Summers, D.A. (1972). Cognitive control. *Psychological Review*, *79*, 58–67.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263.

Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*, 116–131.

Hoffman, R.R. (Ed.). (1992). *The psychology of expertise: Cognitive research and empirical AI.*

New York: Springer-Verlag.
Hogarth, R. (1987). *Judgement and choice* (2nd Edn.). New York: John Wiley & Sons.
Jasanoff, S. (1987). Contested boundaries in policy-relevant science. *Social Studies of Science*, *17*, 195–230.
Jasanoff, S. (1990). *The fifth branch: Science advisors as policymakers*. Cambridge, MA: Harvard University Press.
Jasanoff, S. (1993). What judges should know about the sociology of science. *Judicature*, *77*, 77–82.
Kantrowitz, A. (1967). Proposal for an institution for scientific judgment. *Science*, *156*, 763–764.
Kerr, R.A. (1989). Turning down the heat. *Science*, *246*, 1118–1119.
Kirwan, J.R., Barnes, C.G., Davies, P.G., & Currey, H.L.F. (1988). Analysis of clinical judgment helps to improve agreement in the assessment of rheumatoid arthritis. *Annals of the Rheumatic Diseases*, *47*, 138–143.
Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C.R.B. Joyce (Eds.), *Human judgment: The SJT view* (pp.l15–162). North-Holland: Elsevier.
Koshland, D.E. (1987). Fraud in science. *Science*, *235*, 141.
Ladd, E.C. Jr., & Lipset, S.M. (1972). Politics of academic natural scientists and engineers. *Science*, *176*, 1091–1100.
Lusk, C.M., Stewart, T.R., Hammond; K.R., & Potts, R.J. (1990). Judgment and decision making in dynamic tasks: The case of forecasting the microburst. *Weather and Forecasting*, *5*, 627–639.
Marcus, A.A. (1988). Risk, uncertainty, and scientific judgment. *Minerva*, *26*, 138–152.
Matzkevich, I., & Abramson, B. (1995). Decision analytic networks in artificial intelligence. *Management Science*, *41*(1), 1–22.
Mitchell, R.C. (1984). Rationality and irrationality in the public's perception of nuclear power. In W.R. Freudenberg & E.A. Rosa (Eds.), *Public reactions to nuclear power: Are there critical masses?* (pp.137–179). Boulder, CO: Westview Press.
Mitroff, I.I. (1983). *The subjective side of science: A philosophical inquiry into the psychology of the Apollo moon scientists*. Seaside, CA: Intersystems Publications.
Mumpower, J.L., & Hammond, K.R. (1974). Entangled task-dimensions: An impediment to interpersonal learning. *Organizational Behavior and Human Performance*, *11*(3), 377–389.
Mumpower, J.L., Livingston, S., & Lee, T.J. (1987). Expert judgments on political riskiness. *Journal of Forecasting*, *6*, 51–65.
Murphy, A.H., & Daan, H. (1985). Forecast evaluation. In A.H. Murphy & R.W. Katz (Eds.), *Probability, statistics, and decision making in the atmospheric sciences*, (pp.379–437). Boulder, CO: Westview Press.
National Research Council. (1983). *Risk assessment in the federal government: Managing the process*. Washington, DC: National Academy Press.
National Research Council. (1989). *Improving risk communication.* Washington, DC: National Academy Press.
Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Pool, R. (1992). The third branch of science debuts. *Science*, *256*, 44–47.
Rogers, D.E., & Gellin, B.G. (1991). AIDS and doctors: The real dangers. *New York Times*, July 16.
Ross, J. (1992). [Letter.] *Science*, *257*, 860.
Schneider, S.H. (1989). *Global warming: Are we entering the greenhouse century*? San Francisco, CA: Sierra Club Books.
Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp.11–23). New York: Plenum Press.
Shanteau, J., & Stewart, T.R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes*, *53*, 95–106.

Stevens, W.K. (1989). Skeptics are challenging dire ‘greenhouse’ views. *The New York Times*, *CXXXIX,* December 13.

Stewart, T.R. (1990). A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Weather and Forecasting*, *5*, 661–666.

Stewart, T.R. (1991). Scientists’ uncertainty and disagreement about global climate change: A psychological perspective. *International Journal of Psychology*, *26*, 565–573

Stewart, T.R., Moninger, W.R., Grassia, J., Brady, R.H., & Merrem, F.H. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting*, *4*, 24–34.

Stewart, T.R., Moninger, W.R., Heideman, K.F., & Reagan-Cirincione, P. (1992). Effects of improved information on the components of skill in weather forecasting. *Organizational Behavior and Human Decision Processes*, *53*, 107–134.

Stewart, T.R., Mumpower, J.L., & Reagan-Cirincione, P. (1992). *Scientists’ agreement and disagreement about global climate change: Evidence from surveys* (Research Report). Albany: University at Albany, State University of New York, Center for Policy Research.

Stewart, T.R., West, R.E., Hammond, K.R., & Kreith, F. (1975). Improving human judgment in technology assessment. *Technology Assessment*, 37–43.

Terman, L.M. (1954). Scientists and nonscientists in a group of 800 gifted men. *Psychological Monographs: General and Applied*, *68* (7, Whole No. 378), 1–44.

Ullman, D.G., & Doherty, M.E. (1984). Two determinants of the diagnosis of hyperactivity: The child and the clinician. *Advances in Developmental and Behavioral Pediatrics*, *5*, 167–219.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Wigton, R.S. (1988). Use of linear models to analyze physicians’ decisions. *Medical Decision Making*, *8*(4), 241–252.