PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Automating Assistance for Safety Critical Decisions

J. Fox

Phil. Trans. R. Soc. Lond. B 1990 **327**, 555-567
doi: 10.1098/rstb.1990.0098

Phil. Trans. R. Soc. Lond. B **327**, 555–567 (1990)
Printed in Great Britain

Automating assistance for safety critical decisions

By J. Fox

Imperial Cancer Research Fund Laboratories, Lincoln's Inn Fields, London WC2A 3PX, U.K.

Computer systems are increasingly being introduced to assist in decision making, including hazardous decision making. To ensure effective assistance, decision procedures should be theoretically sound, flexible in operation (particularly in unpredictable environments) and effectively accountable to human supervisors and auditors. Strengths and weaknesses of classical statistical decision models are discussed from these perspectives. It is argued that more can be learned from human decision behaviour than has traditionally been assumed, and this motivates the concept of a symbolic decision procedure (SDP). The SDP is defined, described in terms of first-order (predicate) logic, and its use illustrated in a decision support system for medicine. We point out that the classical numerical decision procedure is a special case of a generalized symbolic procedure, and discuss the potential for rigorous formalization of the latter. We conclude that symbolic decision procedures may meet requirements for assisting human operators in hazardous situations more satisfactorily than classical decision procedures.

Introduction

A decision is a conscious choice between at least two possible courses of action
(Castles *et al.* 1971)

In recent years computer based decision support systems (DSSs) have become increasingly used in fields like medicine and industrial process-control where decisions may have critical consequences for safety. Performance demands may also be severe. For example, action may be needed urgently; available information may be imperfect or incomplete; even the options open to the decision maker may be incompletely worked out. In parallel with the development of DSSs, and in particular the appearance of 'expert systems', concern has been growing about the potential for catastrophic errors created by these systems and, worse, the potential for catastrophes whose causes cannot be established.

Concern for the risks associated with expert systems is now so strong that it has spilled over into public discussion, such as a television broadcast in which American and British practitioners and critics argue the dangers of using expert systems in medical, military and similar applications[†], and there have been calls for restrictions on the deployment of unsupervised or autonomous systems in safety critical situations (*The Boden Report* 1989). These concerns have implications for the design criteria of DSSs and they fall into two main categories:

(*a*) *Performance issues*

1. The decision procedure used by the DSS must perform well (make or recommend good decisions) even in the face of degraded data.

[†] *Electric Avenue*, B.B.C. 1989

2. It should be able to support a wide range of decision types where necessary.

3. Respond flexibly and appropriately should the circumstances requiring a decision change.

(*b*) *Responsibility issues*

1. If decisions lead to errors it must be possible to establish the reasons for those errors.

2. Where it is practical and appropriate provision should be made for a skilled human supervisor to exercise overriding control.

The central requirements for meeting these demands are that we have a comprehensive and sound yet intelligible decision procedure. Theoretical soundness has been the preserve of classical statistical decision theory. However, I argue in what follows that classical procedures may be unacceptably inflexible in the face of unanticipated events, and therefore require close supervision, but are relatively unintelligible to human users. Decision technology has been developing rapidly in recent years, with expert systems trying to address the problems of flexibility and accountability by using techniques from artificial intelligence (AI), but the growth of decision theory has not remained in step, with expert systems frequently being developed in a worryingly ad hoc way (Fox 1988).

This paper suggests an approach to resolving these problems, which takes some inspiration from observations of human decision behaviour but can be formalized by using logical methods (specifically the first-order predicate calculus). First, a brief review of classical statistical decision theory is given.

Classical decision theory

Perhaps the most forthright statement of what should now be regarded as the classical theory of decision making is due to Lindley (1985):

> ... there is essentially only one way to reach a decision sensibly. First, the uncertainties present in the situation must be quantified in terms of values called probabilities. Second, the various consequences of the courses of action must be similarly described in terms of utilities. Third that decision must be taken which is expected – on the basis of the calculated probabilities – to give the greatest utility. The force of 'must', used in three places there, is simply that any deviation from the precepts is liable to lead the decision maker into procedures which are demonstrably absurd ...

There have been innumerable studies using this expected utility (EU) approach in designing DSSS for medical diagnosis and treatment (de Dombal 1972; Schwarz & Griffin 1986), business and politics (see, for example, Hill *et al.* 1979) and many other areas. How does EU theory meet the above requirements? Experience has shown that when Lindley's criteria are satisfied EU procedures frequently perform well and degrade gracefully if the quality or availability of relevant information falls (requirement 1 above).

Unfortunately, however, many practical decisions do not make it easy to meet Lindley's requirements. For example, classical decision procedures require that prior and conditional probability parameters for all decision options and relevant indicators are unambiguous and available, which they are frequently not. The notion of utility is problematic because costs and benefits are subjective, individualistic and difficult to quantify. The validity of measuring utilities of certain outcomes (such as 'quality of life' after surgery) is highly controversial.

Arguably the most serious concern about the numerical framework, however, is that it is a substantially incomplete account of what it means to make a decision. Classical procedures do not specify how we determine that a decision is required, for example, nor how to identify alternative decision options, nor how to select and manage data acquisition and other processes. In short, classical procedures have not been developed to 'respond flexibly and appropriately should circumstances change' (requirement 3).

Human decision makers, on the other hand, clearly are capable of flexibly organizing and reorganizing their actions. (Indeed human decision analysts are routinely required to 'structure' a decision problem before mathematical techniques can be applied.) How do they do this? Human decision behaviour has been widely studied and modelled in psychology (see, for example, Broadbent 1971; Fischoff *et al.* 1981; Kahneman *et al.* 1982). Theoretical accounts of human judgement have often drawn on the classical analysis (see, for example, Hill *et al.* 1979; Hogarth 1980), but as we shall see human decision making may be based on principles that are rather different from those of subjective expected utility (SEU) theory.

Turning to the questions of accountability and audit we can see that classical theory leaves something to be desired here too. Current DSSs are still too primitive to be delegated responsibility for all aspects of decision structuring and decision making. Consequently designers try to keep a human decision maker 'in the loop' in situations involving risk, to monitor and supervise. Here the character of decision theory can be problematic, both because the mathematical terms are unfamiliar to the non-specialist and, even if familiar, the practical implications of sets of numbers may be hard to establish. Consequently this reduces understandability and scope for intervention by the supervisor.

The remainder of this paper is concerned with whether it is possible to develop a new kind of decision theory, which is inspired by human decision making to yield a procedure that is more flexible and intelligible than classical numerical procedures, but which can nevertheless be provided with sound theoretical foundations.

The rationality of human decision making

The basic assumption of classical decision theory is that it deals with rational choice. The decision maker is a rational individual who attempts to maximize the expected value of his or her actions, in the light of well-calibrated estimates of the likelihood of events and the costs and benefits associated with the outcomes of alternative actions. Many studies have shown that human decision makers fall somewhere below this standard of rationality. Reasons for this are not hard to find; our knowledge and use of quantitative parameters can be imprecise; our preferences may be inconsistent; our ability to recall and bear in mind all relevant parameters imperfect, and we are subject to lapses of attention, information overloading and various forms of physiological stress. It is easy to conclude, and often has been, that human decision making is irrational by comparison with normative procedures.

In my view the standard normative criteria of rationality may be too restrictive. I can hardly deny the above observations but analysis of human decision making has concentrated too much on its weaknesses rather than its strengths. As Shanteau (1987) remarks in an analysis of competent or expert decision making '...my emphasis has been on investigating factors which lead to competence in experts, as opposed to the usual emphasis on incompetence'. Shanteau identifies a number of characteristics of expert decision makers. First, he observes that experts know a lot about their field of expertise. They know what is relevant to a specific decision, they know what to attend to in a busy environment, and they know when to make exceptions to general rules. Secondly, and perhaps most compellingly for this discussion, experts know a lot about what they know and they can make decisions about their decisions (Fox 1984). They have good communication skills and abilities to articulate their decision processes. Furthermore, they know which decisions to make, and which not to; they can adapt to changing task conditions, and they are able to find novel solutions to problems.

Whatever their weaknesses, human decision makers have capacities that are needed for flexible adaptation to an unpredictable and dangerous world. The classical model, in contrast, seems to impose fundamental restrictions that make it difficult to match these human capacities.

(*a*) *Restrictions on flexibility*

The expected utility decision rule presupposes that the conditions requiring a decision are predetermined and static. This is quite unacceptable for an autonomous decision maker, which cannot assume that all circumstances have been enumerated in detail before commissioning. Consequently, rationality criteria must include the ability to be rationally flexible, including the ability to: (i) recognize that a decision is needed; (ii) identify the kind of decision it is; (iii) establish a strategy for making it; (iv) formulate the decision options; (v) revise any or all of the above in the light of new information.

A decision system should be capable of autonomously invoking and scheduling these processes as circumstances demand. Classical theory offers no guidance for developing the necessary techniques.

(*b*) *Restrictions on communication*

The availability of an external agent who takes responsibility for the trickier decisions cannot be assumed, but we still expect DSSs to be accountable to their supervisors, and society at large. Therefore, a high level of communication between human supervisors or auditors wishing to examine, and potentially to intervene in, any aspect of the decision process, must be achieved. In general, a decision maker or DSS needs to be able to reflect on the decision procedure, to be able to examine: (i) decision options (what choices exist); (ii) data (the information available that is relevant to a choice); (iii) assumptions (about viability of options, reliability of data etc); (iv) conclusions (in light of data and knowledge of the setting).

Reflective capabilities should extend to the decision process itself, including: (v) the goals of the decision (what is the decision supposed to achieve); (vi) the methods being pursued (what justifies the current strategy); (vii) characteristics of specific procedures (applicability, reliability, completeness etc.).

The decision process should be able to communicate the results of all such reflections for accounting and audit, and should permit intervention in any aspect of the decision process by a supervisor.

A 'rational' theory of decision making must acknowledge these requirements. Classical decision procedures may be optimal in the sense that they promise to maximize the expected benefits to the decision maker, but they must be viewed as unsatisfactory in other ways. The root cause is that the theory fails to provide a framework for fully autonomous decision making, or for audit and sharing responsibility with a human supervisor.

The fact that human decision makers show a degree of flexibility and articulacy, which is dramatically greater than we can achieve with computers that use numerical algorithms alone, suggests ways of improving on current techniques. The next section presents a brief discussion of the origins of these human capacities.

A computational view of human decision making

From a computational point of view EU theory makes a strong commitment to a very specific representational formalism (by using real numbers to encode uncertainty, costs and benefits etc.) and to a process formalism (the algorithmic application of arithmetic operations). From an analogous perspective the representations and processes of human decision making look rather different.

(*a*) *Representation*

Human decision making does not depend much on numbers. Its most prominent features are often said to be an emphasis on recognizing patterns of data that are strongly associated with stereotyped situations (for example, an elderly patient with unexplained weight loss could have cancer) and relations are qualitative and varied (A causes B; A suggests B; B is a side effect of *A*, and so on).

Skilled experts are frequently able to discuss how they make their decisions as well as to make them. For example, they may say what decisions are current, the reasons why they are being attempted, which decisions have been recently taken, what remains to be done, and so forth. We would like to represent the decision process in a way that allows comparable abilities and for this logical and symbolic representations seem promising. Symbolic representations can provide an explicit record of the events, processes, conclusions, justifications etc, from which reports of the decision process can be constructed as required.

(*b*) *Processes*

Classical theory requires the manipulation of numerical probability and uti!ity coefficients whereas human decision making seems to be more strongly influenced by 'heuristic' processes (Tversky & Kahneman 1974). These can be viewed as approximations to a correct decision model, or rules-of-thumb, or pragmatic strategies for rapid decision making. It is widely considered that human decision heuristics only embody probability or utility information implicitly, if at all.

Heuristics can be interpreted as expressing a restricted kind of logical reasoning. For example, decision rules may be expressed in a special case, or propositional, logic; if the patient is elderly and has lost weight then cancer should be considered (Fox 1980). Human reasoning may also exploit 'first-order' rules. These are reminiscent of inference rules in the predicate calculus, which refer to classes rather than individuals, as in:

> If Patient presents with Complaint and possible cause of Complaint is life threatening then investigation of Patient for Complaint is necessary.

Here capitalized terms are logical variables that can be read to mean 'the class of all patients', 'the class of all complaints' etc. (Technically, the variables are universally quantified.)

Whether human problem-solving and decision making truly use propositional or predicate logic (as opposed to processing information in ways which merely resemble logical inference) is an ambiguous and controversial question. For present purposes, however, we merely claim inspiration from human performance without proposing to simulate psychological mechanisms. Just as probability theory provides a clear framework for EU theory, classical logics are formally well-defined, suggest profitable ways of analysing decision processes, and they can be used to clearly specify versatile and intelligible DSSs as shown here.

The oxford system of medicine†

The Oxford System of Medicine (OSM) is an information and decision support system which is intended to provide: (a), a comprehensive medical information service, and (b), assistance to medical practitioners making a wide range of patient management decisions (Glowinsky *et al.* 1989). Among the design objectives for the system are that it should assist with a comprehensive range of decision types (diagnosis, treatment, investigation, risk assessment and referral etc.) while permitting close monitoring and supervision of decision processes. The OSM is aimed principally at medicine, specifically general practice, but the decision process is intended to be application-independent.

Figure 1 shows the way in which a symbolic decision procedure is encoded in the OSM; the decision procedure used employs first-order logic to encode a general decision strategy,

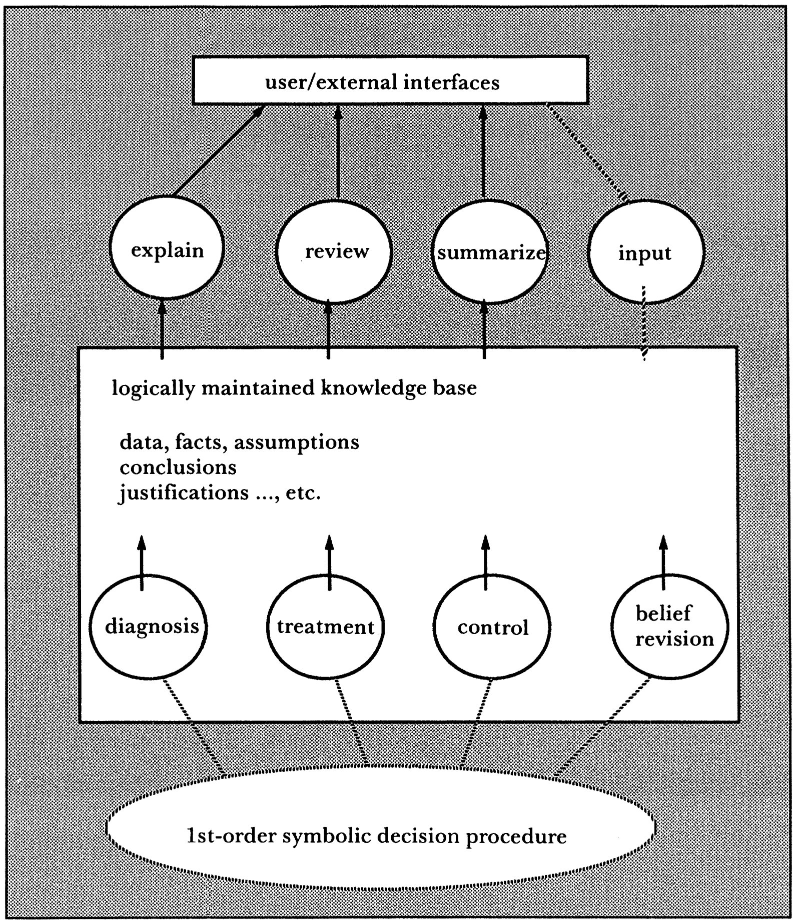

Figure 1. The functional organization of the Oxford System of Medicine. The symbolic decision procedure uses knowledge about specific decision tasks to interpret data input by the user, and to provide explanations, reviews, etc. Technically the knowledge base is logically maintained to automatically ensure consistency of beliefs using a truth maintenance system and logical knowledge of belief states (see text).

† The Oxford System of Medicine and some of the ideas in this paper were developed in collaboration with Andrzej Glowinski and Mike O'Neil. The early work on the design was partly supported by Oxford University Press.

incorporating methods for the following: (i) generating decision options; (ii) reasoning about the pros and cons of the options as data are acquired; (iii) grouping options together which are in some way inter-dependent; (iv) maintaining quantitative preference orderings over the (independent) options; (v) maintaining qualitative evaluative terms for the options.

The OSM incorporates a strategy for creating and executing specific decision tasks in response to user instructions. The details of specific medical decisions (diagnosis, treatment decisions, etc.) are represented as facts or propositions that are consulted by the decision procedure when a specific decision is required. These specify the methods and steps required to satisfy particular decision goals (eg. diagnosis, treatment). Strategies for controlling a decision process are also explicitly defined as part of the system's knowledge.

The abstraction of a general decision procedure from the specifics of medicine (or any other application) and the descriptions of particular kinds of decision and control strategies is central to satisfying some of the requirements discussed above. It contributes to formalization of the theory because essentials are clearly separated from the details of a particular decision; to accountability, similarly, because facilities for explanation and intervention can be designed, which are independent of the circumstances in which they may be required, but having access to all the system's knowledge of the application and knowledge of decision making strategies; to flexibility because the database of facts can be used for multiple purposes and, most importantly, because the system can reflect upon its own operation. The next section elaborates on and illustrates these ideas.

Symbolic decision procedures

'A symbolic decision procedure (SDP) is an explicit representation of the knowledge required to define, organize and make a decision, and is a logical abstraction from the qualitative and quantitative knowledge that is required for any specific application. A SDP may include a specification of when and how the procedure is to be executed.'

A SDP shares some characteristics of human decision making while preserving logical clarity. Human decision making is vulnerable to many influences that affect performance, so our aim is not to emulate human decision making in detail but to achieve a framework which embodies its desirable features within a rigorous computational framework. Features of the current design can be discussed in terms of representation and processing as introduced earlier.

(*a*) *Representation*

The symbolic decision procedure in the OSM represents patient data, medical facts, inferences made during decision making and their justifications, etc. explicitly as propositions, such as:

1. complaint('John Smith', weight-loss) *patient data*
2. causes(weight loss, cancer) *medical fact*
3. kinds(cancer, colon cancer)
4. possible(diagnosis('John Smith', colon cancer)) *a hypothesis*
5. confirmed-arguments('John Smith', diagnosis(colon cancer), support (causes (cancer, weight loss), kinds(cancer, colon cancer)) *a justified argument*

(*b*) *Processes*

These propositions are used by first-order logical inference rules. To put it another way, the rules, and hence the decision procedure, are generalized over classes of concept so that the definition is abstracted from the details of the (medical) application. Just as, say, Bayes' rule for the revision of probabilities specifies how to compute the posterior probability with the set of hypotheses, sources of evidence etc. supplied as numerical parameters to the procedure, a symbolic decision procedure is generalized as a first-order process that accepts qualitative rather than quantitative parameters. First-order processes for a symbolic decision procedure which deal with initiation and control of a decision process, reasoning about decision options, applying numerical decision procedures, achieving flexibility and robustness, and techniques for accountability and audit are now described.

Initiation and control of decision processes

The OSM is principally designed to assist with various medical decisions as required by a doctor. The decision procedures will therefore be initiated by explicit commands. More generally, however, symbolic decision procedures can initiate and control decision processes without supervision. For example, some sort of medical or industrial monitoring system may detect an abnormal parameter, decide to establish its cause and then go on to establish the appropriate treatment or action. If 'rapid weight loss' were observed, for example, and it could be logically established that cancer was among the possible causes of the weight loss (and cancer is of course pathological) then the rule would generate a requirement to take a diagnostic decision (Fox *et al.* (1989) provide more detail.). Recall Shanteau's observation that human experts know which decisions to take; they appear to be able to capture similar capabilities by logical reasoning about decisions and their preconditions.

Reasoning about decision options

Once a decision context has been established a number of basic operations will be carried out in that context, notably to reason about the options, relate options to each other (eg. one may be a special case of another), maintain preference orderings among them, and so forth (Fox *et al.* 1988). For example, we can define logic schemas that *argue* the pros and cons of a decision option. Each takes a number of symbolic terms as parameters, such as what the decision is, the set of decision options that are being considered and the set of arguments identified as supporting them. Argumentation is the process of constructing lines of reasoning and yields reasons for and against decision options for some specific case.

The decision procedure is modelled as a set of general schemas (that is, containing existentially or universally quantified variables) that operate over a database of propositions (see examples 1, 2 and 3 above). Proposition constants with the variables, and case-specific conclusions (such as 3 and 4) are deduced and added to the database. In what follows F, a finding, is any item of case data and O any decision option which might be considered for the case. A_T is an argument based on some theory T, about the truth of proposition P. The braces { } show sets, $\wedge$ is logical conjunction, the right arrow $\rightarrow$ logical implication and $\forall$, $\exists$ are universal and existential quantifiers.

A confirmed argument, conf_arg (C, P, A_T, S), is an assertion that a line of argument A_T about a proposition P applies to the current case, C. In the following scheme for conf_arg the sign, S, qualifies the argued proposition with one of: strengthening $(+)$, weakening $(-)$, confirming $(++)$ or excluding $(--)$.

$$\forall F \forall C \cdot (\mathrm{finding}(C, F) \wedge \mathrm{provable}(F, P, A_T, S) \rightarrow \mathrm{conf\_arg}(C, P, A_T, S)).$$

Where the predicate provable(Db, P, A_T, Q) means 'P may be proved from database Db by argument A_T from theory T with qualifier Q'.

The construction of arguments about decision options are particularly important; where there is a reason for considering the option O and no reason to exclude it then it is added to the option set for the case C, {option (C, O)}.

$$\exists A_{T_i} \cdot ((\mathrm{conf\_arg}(C, O, A_{T_i}, S) \wedge S \in \{+, ++\} \wedge \sim \exists A_{T_j} \cdot (\mathrm{conf\_arg}(C, O, A_{T_j}, --))) \rightarrow \mathrm{option}(C, O)).$$

Here the predicate calculus has the advantage over wholly numerical methods of being able to express methods for introducing (and excluding) decision options incrementally, as information about the decision situation is accumulated.

Flexibility and robustness

The critical, indeed safety critical, question for DSSs is how we can provide them with the capability to take adaptive action in the absence of detailed knowledge of all the circumstances they may encounter. The difficulty with contemporary systems is that the facts or parameters required for all possible situations must be identified in advance. The contrast with people seems obvious; we are not immobilized by ignorance but can fall back on general knowledge of analogous circumstances and general problem solving strategies. A doctor can go back to first principles when faced with a difficult case, or apply general rules of good care while waiting for definitive findings.

One important resource available when facing new situations is theoretical knowledge. Theories can be introduced to interpret, hypothesize and predict circumstances that have not been encountered before because they capture important regularities of the world which (by definition) can be used to make predictions in unfamiliar circumstances. 'Common sense' understanding (of cause and effect; structure and organisation; function and purpose etc.) are included in the term theory. More sophisticated frameworks are also included, such as theories of biological and physical processes (which might be needed in medical decision making) and of course probability provides a well-developed theory for arguing about expectations. If we view theories as first-order inference procedures (that is rules that are generalized over significant classes of events) then theories can be formalized logically and used to argue hypothetically in circumstances where specific (explicit) medical knowledge is incomplete, uncertain or unobtainable, or to augment specific knowledge (see Fox *et al.* 1989) for more detail).

Symbolic procedures may be able to contribute to robustness in other ways. For example, skilled human problem solving and decision making can operate on two levels: the problem solving itself, and a reflective level which monitors and modifies the decision process as information is obtained. New information may suggest additions or corrections to the decision options, cast doubt on the evidence, suggest that the decision or problem needs to be reformulated, and so forth.

Numerical decision procedures as a special case of a SDP

Logical procedures for reasoning about the arguments for and against different decisions carry no information about the weight that should be attached to those arguments. However, although numerical representations may be relatively impoverished, we have acknowledged that classical decision procedures are well understood and they can offer quantitative precision. Symbolic decision procedures do not sacrifice the capability to use them where this is appropriate and practical. The logical elements can be extended with conventional probability and utility calculations. For example, we could include rules that revise probabilities in the light of findings. The schema below illustrates one way of specifying Bayes' rule for updating the probability of a hypothesis from the prior probability of the hypothesis and the finding and the conditional probability of F given H.

$$\forall F \forall H \cdot$$
$$(\text{probability}(H, P_H) \wedge \text{conditional\_probability}(F, H, P_{F|H}) \wedge \text{probability}(F, P_F) \wedge \text{bayes}(P_H, P_{H|F}, P_F, P_{H_{post}}) \rightarrow \text{probability}\ (H, P_{H_{post}})),$$

where $\text{bayes}(P_H, P_{H|F}, P_F, P_{H_{post}})$

is a function which returns a value for $P_{H_{post}}$ and is equivalent to

$$P_{H_{post}} = (P_H * P_{H|F})/P_F.$$

Calculation of expected utility, etc. can also be specified with such schemata (though note that some implementation details have been omitted for clarity).

Returning to the topic of robustness an ability to reflect on the decision procedure also offers advantages here. Mathematical operations, such as Bayesian revision, can be represented symbolically like any other concept. This permits the decision procedure to be extended to include schemata for reasoning about probability revision. Consider the case where the conditional probability table (of symptoms and diseases, say) is incomplete because hypotheses are generated in response to unusual combinations of observations. Under these circumstances the system should not ignore the missing data or, worse, fail to function. It should disable the Bayesian procedure, and continue to operate with, say, qualitative reasoning, while ensuring the problem is communicated to a supervisor or auditor.

Accountability and supervision

I have remarked several times that if decisions lead to errors it must be possible to establish the reasons for those errors, and that provision should be made for a supervisor to exert control. The OSM, for example, provides a range of reporting capabilities summarizing the decision options being considered at any moment, revising the direct and indirect arguments for any option, explaining why specific items of information may be relevant to the current decision and so on.

Good communication requires a rich vocabulary permitting the supervisor to ask questions like 'what diagnoses are presently suspected?', 'why is such and such implausible?' and to assert sentences like 'damage to heart is possible', 'treatment of breathlessness is urgent'. It may be practical to design such a language. A base proposition B such as 'diagnosis of Fred is cancer', can have distinct modalities such as possible *P*; unlikely *P*; suspected *P*. It is useful to compute certain modalities (such as 'possible diagnosis of Fred is cancer') or the general predicate mode(Decision, Case, Option, Mode)) from the pattern of support for the option (O) recorded in the database:

$$\text{let Support} = \{\text{conf\_args}(\mathrm{C}, \mathrm{A}_{\mathrm{T_i}}, \mathrm{O}, \mathrm{S})\} \text{ then}$$
$$\text{option}(\mathrm{C}, \mathrm{O}) \wedge \text{provable}(\text{Support}, \mathrm{A}_{\mathrm{T_j}}, \mathrm{O}, \mathrm{M})\} \rightarrow \text{mode}(\mathrm{D}, \mathrm{C}, \mathrm{O}, \mathrm{M}).$$

Figure 2 shows part of an extended vocabulary for talking about beliefs; the modal terms are defined in terms of logical proof rules (definitions are given in the form of specific modal predicates for clarity).

Terms such as those in figure 2 have often been assumed to have a probablistic interpretation (for example, to say that 'cancer is possible' really means that cancer has a non-zero probability, or 'cancer is probable' that the probability is greater than 0.5 but less than 1.0, or some such.) These numerical interpretations raise many theoretical, empirical and technical problems, and I suspect that a logical interpretation is closer to their linguistic use (Fox 1987). This suspicion has received experimental support recently from an elegant series of experiments

If *arguments* can be identified which support an assertion, *P*, then *P* is *supported*:

arguments for *P* include argument→supported *P*.

P is *possible* if there is at least one supporting argument and *P* cannot be logically eliminated. (NB. This is an *a fortiori* definition. *A* proposition can also be *a priori possible* if the requirement for a supporting argument is relaxed.)

supported *P*, not eliminated *P*→possible *P*.

Arguments are arbitrary proofs over a body of knowledge. We may, for example, argue that a disease is a *possible explanation* of an observation by reasoning from physiology or other theories. Similarly we argue an assertion is *impossible* because it implies that an established theory is violated.

arguments against *P* include violation of Theory → impossible *P* (*and* hence → eliminated *P*).

Arguments can be counted to yield an unweighted measure of *support* (for a diagnosis, treatment, etc.):

possible *P*, number of arguments for *P* = Pos, number of arguments against *P* = Neg → support of *P* is Pos-Neg.

which can be translated into qualitative terms by testing simple arithmetic relations:

possible *P*, support of *P* is *SP*, *SP* > 0, not(support of *Q* is *SQ*, *SQ* > *SP*) → most supported *P*
possible *P*, support of *P* is Support, Support < 0 → dubious *P*.

We can define the concept of plausibility as 'a proposition is *plausible* if it is supported and there are no supporting facts or arguments that are themselves dubious':

supported *P*, not(*F* supports *P*, dubious *F*) → plausible *P*.

and *suspicion* rests on a slightly stronger condition:

plausible *P*, not(possible *P′*, support of *P′* > support of *P*) → suspected *P*.

Similarly, statements which have a conventional probability assignment can be tested to yield qualitative descriptors (strictly a partial ordering):

probability of *P* is Prob, not(probability of *Q* is ProbQ, ProbQ > Prob) → probable *P*
probability of *P* is ProbP, probability of *Q* is ProbQ, ProbQ > ProbP → improbable *P*.

Interestingly, probabilities can also have a second-order logical interpretation, as when we view a high probability as a kind of argument for an uncertain judgement:

probable *P* → arguments for *P* include highest probability of *P*.

So that *probability* arguments can, as intuition requires, yield judgements of *plausibility*.

FIGURE 2. Operators for natural uncertainty terms with logical scheme definitions. *P* is any proposition whose truth is uncertain. ("," is used here to represent logical conjunction.)

with native English speakers by Clark (1989) who was able to show by scaling analyses of some 50 belief terms that the underlying semantic space is multidimensional and highly organized.

The possibility that we can develop generalized logical vocabularies for urgency, importance, doubt and so on, which are closely allied to standard usages, may be of great value for human–computer communication. It should also be noted that a more expressive terminology will have value for decision making itself. For instance the concept of possibility plays a pivotal role in many decision schemata (Fox *et al.* 1989), and terms like plausible, suspected and so on may permit us to express more subtle yet intelligible conditions (Fox 1985). It can also be seen (see bottom of figure 2) that concepts such as 'probable' may be given a logical interpretation, both to express their meaning as a distinct set of logical conditions, and to exploit the concept, in a higher-order inference process. Just as Bayesian decision making can be viewed as a special case of a generalised symbolic decision procedure, so probability can be seen as a specialised uncertainty representation within a logical framework.

Soundness and formalization

One obvious feature of expert systems to date is that they do not have a well-defined decision procedure, but rather a collection of special-case rules for accumulating evidence. Consequently the decision strategy is implicitly distributed over the set of rules, which makes its behaviour uncertain and also a reflective capability is difficult to achieve. Furthermore, expert systems are frequently large and complex pieces of software; it is increasingly recognised that it is difficult to ensure the reliability of such software. Neither of these two characteristics is acceptable for hazardous decision making, but one can argue that the symbolic approach to decision making has potential for improving soundness of DSSs by formal analysis at a number of levels.

1. The semantics of the symbolic terms and propositions stored within the DSS potentially provide strong inter-constraints on the consistency of assertions: a man cannot have a gynaecological condition; symptoms do not cause diseases; a disease is not a possible option in a treatment decision, though it is of course a possible option for a diagnostic decision. These inter-constraints are useful for maintaining the integrity of the system as its knowledge increases, by manual revision or, potentially, by automatic methods of data and knowledge acquisition.

2. Constraints on the initiation and control of a decision derive from knowledge about situations and different types of decision. For instance, it does not make sense to try to make a decision among a number of possible decision options when the action in each case would be the same (as when it makes no sense to discriminate among a set of diagnoses when the treatment would be the same for all of them).

3. The clear definition of specific types of decision (such as diagnosis, treatment etc.) places logical constraints on how a knowledge base should be populated with propositions and theories that are required for each class of decisions.

4. Finally, it should be possible to formally specify a first-order decision procedure, to analyse its abstract properties separate from details of specific applications like medicine, and enforce appropriate class restrictions on variable assignments for example. It may even be possible to establish an a priori understanding of a procedure's performance characteristics, boundary conditions, failure modes and so forth. These are subjects of our further research.

Conclusion

The past few years have seen growing use of computer systems in hazardous settings, in both autonomous and supervized roles. The use of decision support systems, particularly expert systems, is attracting substantial concern. These concerns are well founded, but the growing calls for a moratorium on the introduction of such systems seem unrealistic. A different approach is to develop and formalize a symbolic decision theory that provides the basis for more versatile, more robust, more controllable and more accountable decision support systems than those based on classical expected utility theory, or first-generation expert systems. In this paper I have tried to show that human decision making, for all its vulnerability to error, is a source of inspiration for such a theory, and to illustrate how this inspiration can be turned into practical but formalizable techniques.

I thank Donald Broadbent, F.R.S., Dominic Clark and Mike Sternberg for helpful comments made during the preparation of this paper.

References

Boden, M. (chair) 1989 Benefits and risks of knowledge-based systems. *Report of Council for Science and Society*. Oxford University Press.

Broadbent, D. E. 1971 *Decision and stress*. London: Academic Press.

Castles, F. G., Murray, D. I. & Potter, D. C. (eds) 1971 *Decisions, organisations and society*. Harmondsworth: Penguin.

Clark, 1989 Psychological aspects of uncertainty and their implications for artificial intelligence. Ph.D. thesis, University of Wales Institute of Science and Technology.

de Dombal, T. 1979 Computers and the surgeon: a matter of decision. *Surg. Ann.* **11**, 33–57.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S. L. & Keeney, R. L. 1981 *Acceptable risk*. Cambridge University Press.

Fox, J. 1980 Making decisions under the influence of memory. *Psychol. Rev.* **87**, 190–211.

Fox, J. 1984 Formal and knowledge-based methods in decision technology *Acta. Psychologica*, **56**, 33–331. (Reprinted in 1988 *Professional judgement* (ed. J. Dowie & A. Eldstein). Cambridge University Press.

Fox, J. 1987 Making decisions under the influence of knowledge. In *Modelling cognition* (ed. P. Morris). London: J. Wiley.

Fox, J. 1989 Symbolic decision procedures for knowledge based systems. In *Handbook of knowledge engineering* (ed. H. Adeli). New York: Prentice-Hall.

Fox, J., Glowinski, A. & O'Neil, M. 1987 Towards a knowledge based information system for primary care. In *Information handling in general practice* (ed. R. H. Westcott & R. Jones). London: Croom Helm.

Fox, J., Glowinski, A., O'Neil, M. & Clark, D. A. 1988 Decision making from a logical point of view. In *Research and development in expert systems V* (ed. B. Kelly & A. Rector), pp. 160–175. Cambridge University Press.

Fox, J., Clark, D. A., Glowinski, A. J. & O'Neil, M. 1990 Using first-order logic to integrate qualitative reasoning and decision theory. *IEEE Trans. Syst. Man Cybern.* (In the press.)

Glowinski, A. J., O'Neil, M. & Fox, J. 1989 Design of a generic information system and its application to primary care. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine, Lecture Notes in Medical Informatics* (ed. J. Hunter), pp. 221–233. Berlin: Springer–Verlag, 1989.

Hill, P. H., Bedau, H. A., Chechile, R. A., Crochetiere, W. J., Kellerman, B. L., Ounjian, D., Pauker, S. G., Pauker, S. P. & Rubin, J. Z. 1979 *Making decisions*. Reading, Massachusetts: Addison-Wesley.

Hogarth, R. M. 1980 *Judgement and choice: the psychology of decision*. Chichester: John Wiley.

Kahneman, D., Slovic, P. & Tversky, A. (eds) *Judgement under uncertainty: heuristics and biases*. Cambridge University Press.

Lindley, D. V. 1985 *Making decisions*, 2nd edn. London: John Wiley.

Schwartz, S. & Griffin, T. 1986 *Medical thinking*. New York: Springer.

Shanteau, J. 1987 Psychological characteristics of expert decision makers. In *Expert judgement and expert systems NATO ASI Series* (ed. J. Mumpower), vol. F35.

Tversky, A. & Kahneman, D. 1974 Judgement under uncertainty: heuristics and biases. *Science, Wash.* **185**, 1124–1131.