The Knowledge Engineering Review, Vol. 5: 1, 1990, 1–4

Safe expert systems: simulating experts or building formal theories?

JOHN FOX

This issue of the *Review* includes two papers dealing with decision theory. 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 example, outcomes may have consequences for safety and action but may be needed urgently though available information may be imperfect or incomplete and 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 therefore 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 British television programme[1] in which American and British practitioners and critics argued the dangers of using expert systems in medical, industrial, military and other applications, and there have been calls for restrictions on the deployment of unsupervised or autonomous systems in safety critical situations (The Boden Report, 1989).

Decision making is pivotal to AI generally, not just expert systems. Systems for natural language understanding must decide between alternative interpretations of sentences; vision systems must resolve ambiguities in images. Problem solvers, planners and automated design involves deciding between alternative solutions, actions, or device components, taking into account the uncertainty arising from the unreliability and/or imprecision of available information. If significant progress in these areas leads to practical technologies then no doubt such technologies will all find their way into safety critical applications.

The underlying problem, addressed in different ways by the paper on behavioural decision theory from Lehner and Adelman and on decision theory and planning from Haddawy and Rendell, is that research on sound decision procedures has not been given sufficient emphasis in either AI or knowledge engineering research.

A great deal of theoretical work in developing formal foundations of AI, for example, has concentrated on qualitative reasoning and logical inference because these are seen as novel and promising foundations for building sophisticated automata. Consideration of the practical circumstances in which symbolic reasoning systems may be used has led to developments in defeasible reasoning, such as non-monotonic logic and truth-maintenance systems, that take account of the dynamic acquisition and use of unreliable information. Yet, in any practical application of AI, inference is only part of the story. Equally important is the decision-theoretic framework within which inference systems operate, which in the end determines the actions that should be taken by the decision maker. This must not only take into account whether inferences can be made soundly or not but also the cost utility of possible actions, how much information is needed to safety select them and what should happen when more information comes in and beliefs about decision outcomes are questioned.

The paper by Haddawy and Rendell reviews a particular field of AI that has been largely concerned with qualitative reasoning – planning. While recognizing the technical sophistication that

[1] Electric Avenue, BBC, 1989.

the field has achieved, and the growing importance of formal logic, they argue that much more is needed and that classical theories of how decisions under uncertainty *ought* to be made has a significant contribution. The elements of classical decision theory are widely known but principled extensions to the theory to cover the new tasks addressed in AI, such as planning and design, time-dependent reasoning, meta-level inference and so forth, raise important challenges which range well beyond the authors' focus.

A prominent idea in practical knowledge engineering and research has been that a (perhaps *the*) source of inspiration for system design is what human experts do. Taking this approach in designing DSSs we must elicit answers from experts to such questions as "what are the faults most likely to be seen on such and such a device?", "how likely is it that such and such a component will fail in such and such a way?". Lehner and Adelman draw attention to the large body of research into human judgement and decision making, carried out over several decades, which is not well known in the knowledge engineering community. Moreover they point out that this research has yielded a good knowledge of common biases and weaknesses in human judgement, particularly probabilistic judgement. Reasons for these weaknesses are not hard to find; our knowledge and use of quantiative parameters can be imprecise; our preferences may be inconsistent; our ability to recall and bear in mind all relevant parameters imperfect, and even experts are subject to information overloading and lapses of memory and attention. The accumulation of knowledge about human decision capabilities should be considered carefully before taking "expert knowledge" at face value. The authors provide a number of useful references for those wanting to follow up this subject.

It seems clear that a good knowledge of both statistical decision theory and behavioural decision theory should be a part of the basic training of knowledge engineers and researchers in applied AI, and that there may be important gaps to be filled in both the curriculum and research agenda. However achieving a sound hybrid of knowledge-based and decision-theoretic techniques is not all that is needed if we are to build sophisticated DSSs. A hint about something else that is missing can be found in the psychological literature.

The basic assumption of decision theory is that it deals with *rational choice*. The decision maker is a rational agent which (or who) attempts to maximize the expected value of its 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. As Lehner and Adelman discuss, many studies have indicated that human decision makers fall somewhere below this standard of rationality. It is easy to conclude, and often has been, that human decision making is irrational by comparison with formal statistical decision procedures.

However the statistical criteria of rationality may be too restrictive. One can hardly deny the above observations about human performance but analysis of human decision making has tended to concentrate on its weaknesses rather than its strengths. As Shanteau (1987) remarks in an analysis of 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 positive characteristics of expert decision makers. Firstly, 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, experts *know a lot about what they know* and they can *make decisions about their decisions*. 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.

Such observations raise additional issues for the designers of DSSs, whether these are statistical, knowledge based or hybrid systems. They fall into two main categories, performance and responsibility.

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. Robustness entails being able to assess the reliability of information sources and to seek alternatives where necessary, as well as merely cope with uncertainty.

(2) Few practical situations involve just one class of decision (e.g. diagnosis); decision theory must surely address the problem of deciding what decision is required.

(3) Many practical automata must face rapidly changing situations, not only in the information available but also in the problem that needs to be solved. Decision support systems must incorporate capabilities for altering their decision goals as circumstances develop.

The central requirements for meeting these demands are that we have a sound yet flexible decision procedure. To date theoretical soundness has been the preserve of classical statistical decision procedures. However classical procedures are inflexible in the face of unanticipated events, and therefore may require critical monitoring and supervision. DSS assessment criteria must include the ability to be *rationally flexible*, including the ability to.

a. recognize that a decision is needed
b. identify the kind of decision it is
c. establish a strategy for making it
d. formulate the decision options
e. 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 little guidance for developing the necessary techniques.

Responsibility issues

We must also achieve a high level of communication between human supervisors and/or auditors wishing to examine, and potentially to intervene in, any aspect of the decision process.

(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 supervisor to exercise overriding control.

In general a decision maker needs to be able to *reflect* on the decision procedure, to be able to examine the:

f. decision options (what choices exist)
g. data (the information available that is potentially relevant to a choice)
h. assumptions (about viability of options, reliability of data etc.)
i. conclusions (in light of data and knowledge of the setting)

Reflective capabilities should extend to the decision process itself, including

j. the goals of the decision (what is the decision supposed to achieve)
k. the methods being pursued (what justifies the current strategy)
l. characteristics of specific procedures (applicability conditions, reliability, completeness etc.)

A theory of rational decision making must asknowledge these requirements. Classical decision procedures may be optimal in the sense that they promise to maximise the expected benefits to the decision maker, but they must be viewed as unsatisfactory in other ways. It seems to me that the root cause is that even a clever combination of logical and probabilistic inference is not by itself an adequate basis for fully autonomous decision making (Fox, Clark, Glowinski and O'Neil, 1990).

The papers in this issue rightly emphasize the importance of statistical decision theory and awareness of relevant research. An additional challenge for the future, however, may be to find ways of understanding how good human decision makers achieve what they achieve, to try to improve on

these capacities using sound theories of logical and statistical reasoning, and to develop a theory of decision making which meets the need for reflection.

References

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

Fox, J, Clark, DA, Glowinski, AJ and O'Neil, M, 1990. "Using predicate logic to integrate qualitative reasoning and classical decision theory." *IEEE Trans. on Systems, Man and Cybernetics* (In press).

Shanteau, J, 1987. "Psychological characteristics of expert decision makers" in J Mumpower (ed) *Expert Judgement and Expert Systems*, NATO ASI Series, vol. F35.