Validation of Expert Systems: Two Perspectives

by

E. Subrahmanian, J. Davis

EDRC-05-11-87

Validation of Expert Systems: Two Perspectives

Eswaran Subrahmanian
Engineering Design Research Center
Carnegie Mellon University
Pittsburgh, PA 15213
Ph.no. (412)-268-2257
Arpanet: eswaran@h.cs.cmu.edu
Bitnet: es3e%te.cc.cmu.edu@cmuccvma.bitnet
and
Joseph G. Davis
Department Operations and Systems Management
Graduate School of Business
Indiana University
Bloomington, Indiana 47405
Ph. no. (812)-335-8449
Bitnet: davis1@iubacs.bitnet

12 January 1987

Keywords: expert systems, validation techniques, Artificial intelligence.DSS, Simulation

Submitted to the Seventh International Workshop on Expert Systems and Their Applications,
May 13-15,1987, Avignon, France.

ABSTRACT

Expert systems have generated a lot of interest in both academia and industry. The development of tools, techniques, and methodologies drawn from from artificial intelligence (AI) for building expert systems have not been matched by proposals for evaluating and assesing them. This paper examines the assessment problem in the context of two distinct perspectives of expert systems in the spectrum of possibilities a) expert systems as psychological theories of human problem solving, and b) expert systems as decision aids. We review the assessment methods/approaches for each of the two categories, with the objective of ensuring that they are consistent with the basic assumptions of each category. Based on the law of requisite variety, for evaluation to be effective, the variety of methods employed have to match the variety to be found in the phenomena to be evaluated.

Validation of Expert Systems

1 Introduction

The widespread surge of interest both in academia and industry over the last decade in expert systems (also known as Knowledge based systems - KBS) based on techniques, and methodologies drawn from artificial intelligence (AI) has not been matched by proposals for evaluating or assesing such systems in particular and underlying research in this field in general. Measurement of effectiveness is a difficult problem in the best of times; it is rendered more so in AI by the multiplicity of competing approaches, theories and orientations and the relatively early stage of its development. Bundy has described the pre-paradigmatic state of AI research and some of the informal criteria that researchers (and reviewers) employ for assesing AI research [6].

This paper examines the assessment problem in the context of two distinct perspectives of expert systems in the spectrum of possibilities:

1. expert systems as psychological theories of human problem solving, and
2. expert systems as decision aids.

In an ideological sense, systems should combine both these aspects. However, in practice this dichotomy has been found to exist, primarily due to the influence of cognitive psychology/science on the former and the relatively engineeristic approach on the latter [16]. We review the assessment methods/approaches for each of the two categories, with the objective of ensuring that they are consistent with the basic assumptions of each category. Based on the law of requisite variety, for evaluation to be effective, the variety of methods have to match the variety to be found in the being evaluated.

2 Expert Systems as Psychological Theories

Expert systems are computational models that exhibit behavior exhibited by human experts. In these systems, simulation of human behavior in complex problem solving tasks is done by creating a computational model of the problem task by use of verbal reports collected during the performance of the task. In defining the scope of verbal reports as data Simon and Ericsson state [8]:

> We will conceive of recorded verbalizations as data - exactly like latencies, eye fixation, sequence of moves , and so on - to be accounted for by a corresponding model which generate them literally or on the level of encoded patterns or information content.

They further state that models that can regenerate verbalizations (or encoded aspects of them) can be constructed and tested without making any assumptions about the internal structure of processes. This method is significantly different from psychology which looks for systematic patterns in large bodies of data. An explanation that fits the data is kept till additional studies and data provide counter-evidence to replace the existing explanation with a new one. Many AI researchers, argue that the model used by psychologists are incomplete and trivial as they in their studies ignore a large number of factors that interplay in the study of a mechanism such as word recognition; the studies in isolation do not provide any insight into human behavior which are complex processes that involve interaction between processes that are raised and even contradictory [1]. On the other hand, psychologists feel that AI research is methodologically sloppy as most AI models are restrictive and lack any capability of generalization into a theory.

Sharkey and Pfeifer argue that much of the conflict between cognitive psychologists and AI researchers arise from the fact that the two fields section cognition differently [27]. They claim that psychologists section it horizontally while artificial intelligence researchers section it vertically. In other words, the psychologists, while breaking down the problem of cognition into decomposable parts investigate each part for psychological validity while the AI researcher is more interested in identifying the interaction between these models. It is this difference that is at the core of the differences between psychologists and AI researchers; because an AI researcher, even if not able to demonstrate the working of the mechanism beyond a small set of examples, claims generality due to the complexity of interactions while a psychologist claims generality based on applicability of his model for a wide variety of experimental situations.

AI researchers would say that even though AI has the same subject matter as cognitive psychology, it is subject to a different acceptability criteria [13]. The most predominant criteria for acceptability is that of implementability [6]. Hayes [13], in identifying imptementability as the criteria says, "An acceptable piece of behavior must be in the last analysis, a program which can actually be implemented and run. And such an explanation is a good explanation just to the extent that the program, when run, does indeed exhibit the behavior which was to be explained." In other words, in AI, the only explanation of behavior that is acceptable is when a theory is implemented as a computer program, exhibits the same behavior that the psychological theory explains. This leads to two questions, a) Is implementation a sufficient criteria for validation? and b) What is the performance criteria that is based on the exhibition of behavior to be explained?, The question of implementation as a sufficient criteria has been challenged by Sharkey and Pfeifer on the grounds of two practices within AI research [27]. They observe that most AI programs have what are known as "Kludges". A Kludge is a term used for patches of program used to connect mismatched components of a computer system. In terms of theory and implementation, kludges represent the aspects of mismatch between the two. The other problem they identify is the use of "wishful pneumonics" after McDermott [15]. Here, McDermott refers to the use of terms such as "understand" to name functions that actually correspond possibly to a simple aspect of the understanding process. McDermott claims that the use of these pneumonics are dangerous as they tend to deceive not only the reader but also the programmer himself. Understandably, these practices question the implementation criteria as a sufficient proof to declare an AI program as a physical representation of a theory. This however, does not imply that implementation is not a necessary part of the validation process. In fact, Sharkey and Pfeifer [27] suggest, documentation of kludges between the techniques will go a long way in identifying and modifying the implementation to conform to the proposed theory in subsequent attempts.

The second question is that of identifying the performance criteria for the system based on the exhibition of behavior to be explained. The most commonly talked about performance criteria is the "Turing Test". The Turing Test states that if a human being were to interact with a mechanism through a teletype and the human is not able to differentiate the mechanism from another human being, then the mechanism has passed the test. There is a lot of debate as to what exactly this implies and even in in its weakest form most AI systems currently existing do not pass this test. Leaving aside this strong criteria, a weaker criteria is that of Hayes which states the exhibition of behavior by the program is limited to the behavior that has to be explained. In arguing against this criteria, Sharkey and Pfeifer [27]point out that it is insufficient by illustrating with an example that it is possible to deceive the user by employing transformations of data that are not revealed to the user. This was observed by Weizenbaum in his experience with the users of ELIZA which used keywords in generating subsequents responses [28].

In the previous paragraphs, we have taken the different acceptability criteria for expert systems as psychological theories and presented arguments that these are not sufficient criteria individually. Sharkey and Brown [26] in their argument for the need for empirical foundation of AI claim, "that cognitive science must consist of three interacting parts which continually feed back to one another. These are:

- the construction of theory by whatever means available, e.g. intuition, imagination, formal reasoning, knowledge of people, knowledge of psychological evidence;
- Theory evaluation by empirical testing using objective scientific methods; and
- careful computer implementation to ensure that the theory really is a possible explanation of human cognition.

Some AI researchers view, the task of building systems differently. They believe that artificial intelligence is to psychology and computer science what applied mathematics is to mathematics and physics. Hence, feel that the purpose of AI research, is to identify new computational techniques for cognitive modelling [6]. In their exchange, Bundy and Ohlsson, on the nature of artificial intelligence argue for identification of principles that relate different techniques and their range of applicability as important aspects [2,3,4,5] [19,20,21, 22]. This approach to AI is closer to technological applicability than psychological validity. This leads us to my next question of what constitutes a sufficient criteria if the purpose of the systems is not one of providing a psychological theory but to serve as a decision aid. In the next section I will explore this issue.

3 Expert Systems as Decision Aids

In order to examine the role of expert systems as decision aids, I briefly survey the validation procedures used for different computational models commonly in use as decision aids. Two computational models that are used as decision aids are operations research models and simulation models. Operations research models as decision aids have their objective precisely defined in terms of the solution characteristics. This allows the operations researcher to identify computational methods (algorithms) and specify their performance criteria in terms of space-time complexity. This aspect of operations research is directly transferable to AI systems that primarily use "weak methods." This is evidenced in the study of three different heuristic algorithms for their performance empirically [11]. "Weak methods" can be subjected to performance criteria based on specific measurements mainly because of their domain independence. On the other hand, "expert systems" have been said to use "strong methods"--implying that the system relies heavily on domain dependent knowledge in conducting the search [9]. This, however, does not mean that expert systems do not have the problem of arriving at a solution in a reasonable amount of time but only that they cannot be subjected to the same performance criteria as other search models that are domain independent.

The second type of computational models commonly used as a decision aid are simulation models. Simulation models are more like expert systems in that they are domain and problem dependent while they employ different mathematical techniques to model real systems. The difference between the two lies more in the set of techniques used in the process of simulation; expert systems use predicate logic, heuristics, and evidential reasoning mechanisms to simulate aspects of human behavior in complex problem solving tasks; traditional simulation relies heavily on the mathematical methods such as queueing theory, differential equations and automata theory in describing real world systems in terms of discrete-event-models, discrete-time models, and continuous models [17, 29]. Expert systems and simulation models are different in another dimension as well in that expert systems use data driven programming as the main computational model while simulation models are procedural in nature.

Feigenbaum calls building expert systems "knowledge engineering" and refers to it as the applied side of AI that involves issues of knowledge representation (choice of appropriate data structures), utilization (choice of designs for inference engine) and acquisition (accumulation of new knowledge) [10]. These issues are not new from the point of view of simulation except for the acquisition part. The use of both mathematical and statistical techniques (inference) and data structures that involve symbolic data is another dimension that traditional simulation lacks due to its numeric nature. However, the need for providing complex data structures to model real world objects specific for the use with some mathematical techniques led to the development of specialized languages for simulation. One computational technique commonly used in artificial intelligence, object oriented programming has its roots in SIMULA, a simulation language [12]. From this perspective, expert systems whose purpose are that of decision aids should be termed as knowledge-based systems as the burden of validity of these systems is not based on theorizing in psychology but in their operational and conceptual validity for the purpose the system is to serve. In both expert systems research and in simulation implementation is a necessary criteria for validation. Having identified the similarities and differences between the two different modelling mechanisms, we briefly review the applicability of other validation procedures used in the field of computer simulation for expert systems.

In his survey of verification and validation of simulation models, Sargent defines verification and validation as follows [25]:

> Verification is usually defined as ensuring that the model behaves (runs) as intended, and validation is usually defined as determining that an adequate agreement exists between the entity being modelled and the model for its intended use.

Verification is a topic that has a wide implication for all computer models. The field of software engineering is devoted to design, test, and prove programs correct. While the applicability of these verification methods to expert systems is undeniable, but is beyond the scope of this paper. Hence, I will restrict myself to validation of techniques used in simulation models that are applicable to expert systems. Sargent [25], in his paper, also acknowledges that no procedures exist to guide the choice of techniques for validation, and in practice, they are problem dependent. In choosing the techniques for applicability to expert systems, several of the techniques that involve testing predictive powers of the model have been excluded. The techniques chosen are: a) face-validity, b) traces, c) multistage validation, d) historical data validation, and e) event-validity. I will discuss each of them using Sargent's definitions, in the following paragraphs [25].

Face-validity: Face-validity involves asking experts whether the model is reasonable, i.e. to ask experts whether the results of the computer generated outputs and their internal behavior are reasonable. Face-validity is a useful procedure to identify how the expert may have done differently or to confirm different parts of the reasoning behavior in order to validate the conceptual model of the system.

Traces: In simulation, traces involve studying the behavior of different entities in the model to confirm the correctness of the model. Entities in simulation usually are variables and behavior is the pattern of values taken by those entities during the course of the simulation. In expert systems, one set of entities would involve information on such things as decision sets at different decision making points. The other set of entities would be the set of reasoning mechanisms that are used in the model. Traces of the situations where each of these reasoning mechanisms are used should be evaluated for their correctness and completeness.

Multi-stage validations: Naylor's multi-stage validation consists of [25]: "a) developing the model's assumptions on theory, observation, general knowledge, and intuition; b) validating the model's assumptions where possible, and c) comparing the input-output relationship of the model to the real system." The above multi-stage procedure is similar to the outline of methodology for AI presented in the previous section.

Historical data-validation: This technique involves the use of historical-data to test that the model behaves as expected. In medical diagnosis, medical case histories published by the American Medical Association, has been used in evaluation of performance of diagnostic systems (MYCIN, MDX, and Internist-1), in terms of number of successful diagnosis for a given set of cases [24,7]. Such data may or may not be available easily for many domains but it will be useful to collect such data for testing expert systems.

Event-validitv: This involves testing of concurrence of occurrence of events in a simulation of that of a real system. In the case of expert systems, events would correspond to the information acquisition events. Coherence of a dialogue between expert systems and the user would serve as a test for a valid ordering of events. In the context of event-validity, comparison of problem-behavior-graph could be an acceptable alternative as it corresponds to a flow chart of problem-solution events.

Expert systems, similar to simulation systems, are built for operational use more than for validating a psychological theory. Thus the concept of operational validity is useful and the techniques for operational validity could include: a) historical-data-validation for establishing the reliability of the model in terms of arriving at the correct solution, b) event-validity is important if any of these systems are to be used for pedagogical purposes and also for ease of user interaction, and c) traces of the system's behavior are useful in the form of "explanations" or "regenerated verbalizations" of the problem solving process in order that the user can identify the logic of generation of the solution to evaluate the conceptualization provided by the system. This technique can also be used for conceptual validation.

3.1 User Assessment of Expert Systems as Decision Aids

Probably the least rigorous of the evaluation methods, the objective is to measure users' perceptions of the degree to which the system(s) enable them to make effective decisions in the task domain of interest. The output of this approach is a quantified score of user satisfaction with the system. This method has been applied frequently in research on design, development and, implementation of decision support systems which are decision aids that rely on a database that feeds relevant data.

Measurement of user perceptions and attitudes relies heavily on research in Psychometrics[18]. It typically involves the development of psychometrically valid instruments for measuring the interest, e.g., degree of user satisfaction with an expert system. The sequence of steps involved include:

1. Identification of critical dimensions relating to users' satisfaction with the system in terms of enhancing the quality of decisions based on the literature review and interviews with the users.
2. Testing the instrument for establishing its reliability and validity.
3. Repeated use of refinement over time in a multiplicity of settings.

A few such instruments in the DSS literature are available [23,14]. While the direct relevance of these instruments to expert systems is extremely limited, the approach deserves greater attention given its potential for tracking managerial acceptance and user perceptions.

4 Conclusions

In this paper we have presented a set of methods/approaches that are applicable in evaluating expert systems as psychological theories and as decision aids. This is in no way, an exhaustive listing of methods that are applicable in evaluating expert systems. We hope that in presenting the possible methods/approaches we can look forward to generating results that may lead to a possible prescriptive criteria on the applicability of these methods for different purposes to which expert systems are built to serve.

References

[I] Anderson, J. A.
ArchitectureofCognition.
Harvard University Press, Cambridge, 1981.

[2] Bundy, A.
Superficial Principles: An analysis of a Behavioral Law.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Winter):20-22,1983-84.

[3] Bundy, A.
Nature of AI: Reply To Ohlsson.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Summer):24-25,1983.

[4] Bundy, A.
Principled Behaviour: Bundy vs Ohlsson, Round 4.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(autumn):26-27,1983.

[5] Bundy, A.
Superficial Principles: An Analysis of A behavioural Law.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(autumn):26-27,1983.

[6] Bundy, A.
Some Suggested Criteria for Assesing AI Research.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Spring-Summer)26-28, 1981.

[7] Sticklen, J., Chandershekaran, B., Smith, J. W., and Svirbely, J.
A comparison of Diagnostic Subsystems of MDX and MYCIN.
In *Proceedings of the IEEE Workshop on Principles of Knowledge Based Systems,* pages 205-212. 1984.

[8] Ericsson, K. A., Simon, H. A.
Verbal Reports as Data.
Psychological Review 87(3):215-251, May, 1980.

[9] Ernst, G. W., Banerji, R. B.
On the Relationship Between Strong and Weak Problem Solvers.
AI Magazine Summer:25-29,1983.

[10] Feigenbaum, E.
KnowledgeEngineering:AppliedSideofArtificialIntelligence.
Technical Report, Dept. of Computer Science, Stanford University, August, 1980.

[II] Gashnig, J.
PerformanceMeasurementandAnalysisofCertainSearchAlgorithms.
PhD thesis, Carnegie Mellon University, 1979.

[12] Goldberg, A., Robson, D.
SmallTalk:TheLanguageanditsImplementation.
Addson Wesley, Reading,MA, 1983.

[13] Hayes, P.
On the Difference Between Psychology and AI.
ArtificialIntelligence:HumanEffects.
Ellis Horwood, Chichester, U.K, 1984, pages 157-162.

[14] Jenkins, A. M., and Ricketts J. A.
ThedevelopmentofaDSSSatisfactionQuestionaire.
Artificial intelligence Laboratory, Memo 296, MIT, 1985.

[15] Mcdermott, D., V.
Artificial Intelligence Meets Natural Stupidity.
MindDesign.
MIT Press, Cambridge, Mass, 1980, pages 163-172.

[16] Meltzer.B.
Artificial intelligence.
ArtificialIntelligence and Simulation ofBehaviour Quarterly (47)21-23$_1$1983.

[17] Nilsson, N.J.
PrinciplesofArtificialIntelligence.
Tioga, Palo Alto, CA, 1980.

[18] NunnaHy.
PsychometricTheory.
McGraw Hill Book Co., 1978.

[19] Ohlsson, S.
Tell Me Your Problems: A Psychologist Visits AAAI82.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Winter):27-29,1982-83.

[20] Ohlsson, S.
Mathematics, Behaviour, and Creativity: Reply to Bundy.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(summer)25,1983.

[21] Ohlsson, S.
Mechanisms, Behaviours, Principles: A Time for Examples.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Autumn):24-25,1983.

[22] Ohlsson, S.
A.I. Principles: Functions of Scientific Laws.
ArtificialIntelligenceandSimulationofBehaviourQuarterly(Spring-Summer):36, 1984.

[23] Pearson, S.
MeasurementofComputerUserSatisfaction.
PhD thesis, Arizona State University, 1977.

[24] Miller R., Myers J. D., Pople H.
Dialog - A Computer system for Medical Consultation.
NewEnglandJournalofMedicine ():, 1974.

[25] Sargent, R., G.
Verification and Validation of Simulation Models.
ProgressinSimulationandModelling.
Academic Press, New York, 1982, pages 159-169.

[26] Sharkey, N. and Brown, G.
Why Artificial Intelligence Needs an Empirical Foundation.
ArtificialIntelligence:PrinciplesandApplications.
Chapman Hall, New York, 1985, pages 260-291.

[27] Sharkey, N. and Peifer, R.
Uncomfortable Bedfellows: Cognitive Psychology and Artificial Intellgence.
Artificial Intelligence: Human Effects.
Ellis Horwood, Chichester, U.K., 1984, pages 163-172.

[28] Weizenbaum, J.
Computer Power and Human reason.
Freeman Press, New York, 1976.

[29] Zeigler, B. P.
Theory of Simulation and Modelling.
John-Wiley, New York, 1976.