Making a mind: a cognitive engineering approach

John Fox
Advanced Computation Lab
Imperial Cancer Research Fund
Lincoln's Inn Fields
London UK
jf@acl.icnet.uk

Abstract

In recent years empirical research into AI models of mind have has fallen off in favour of formal and engineering studies of intelligent systems. We argue the need for more "scientific" investigations of mind, with a strong empirical as well as theoretical methodology, and advocate a unified research programme combining knowledge level and architectural theories.

1 Introduction

Before the Japanese 5th, Generation Computer Project stimulated an explosion of interest in the 1980s the AI research community was quite small and focused. Although the field was clearly interdisciplinary, the computer scientists, psychologists, philosophers and others involved had enough unity of purpose to agree to call themselves "cognitive scientists", sharing a common interest in understanding the nature of intelligence, both natural and artificial.

A direct effect of the 5th Generation Project was to draw researchers into AI from many fields that did not have a traditional interest in cognitive science. These included statisticians, operations researchers and software (and other kinds of) engineers. They brought many different perspectives and techniques to the field, but a consequence was that it became fragmented into competing schools of thought. Debates raged, including those of symbolicists vs connectionists, logicists vs Bayesians and experimentalists vs formalists. Indeed a fundamental philosophical split also emerged, between those communities who emphasised experimental *science* (e.g. psychology and neuro-science), *practical engineering* (e.g. knowledge systems and robotics) and *formal theory* (e.g. mathematical logic and bayesian inference). Although one may argue that the diversity of approach was healthy it seems to have had some negative effects, and in particular some loss of scientific consensus about what the field of AI was about.

For reasons which are unclear, but we hope because of a weariness with unproductive factionalism, AI may now be trying to recover the scientific unity which it lost in the nineteen eighties. Proposals for "unified theories of cognition", "intelligent agents", "theories of consciousness" – and this conference – suggest a desire to return to the intellectual roots of the subject; to understand the principles of mind(s).

We are experimental scientists (though with strong interests in practical engineering and formal theories in Al) so we have always had a strong interest in empirical approaches to cognitive science. We have had a particular interest in the experimental and theoretical programme advocated by the late Allen Newell, which was aimed at constructing a "unified theory of cognition" (UTC; Newell, 1990). Frustrated with psychology's tendency to fragment into niche areas like memory, reasoning, decision-making, perception, language and so forth, Newell argued that the time is right to exploit computational ideas to bring these together in understanding complete systems. Terms like "mind" were unfashionable at the time but understanding and building minds was certainly what Newell was about.

2 Understanding and building minds: the SOAR programme

Newell's concept of a unified theory is to be found in its most developed form in the SOAR theory of the human "information processing architecture". SOAR grew out of a long collaboration on human problem solving with Herbert Simon (Newell and Simon, 1972), which stimulated many conceptual and technical developments in cognitive science. The clearest instance of this was the realisation that a very simple computational element, the humble production rule, was sufficient to implement a powerful and important family of symbol-manipulating systems. SOAR built on the successes of earlier production rule technologies like PSG and OPS, by making a couple of small but important extensions to the elementary recognise-act cycle of rule interpretation. The main extensions were the ability to generate *goals* dynamically and to *learn* new rules based on experience. Laird, Newell and Rosenbloom (1987) argued that the resulting computational architecture had the necessary and sufficient means to demonstrate "general intelligence". SOAR was presented as an architecture for general intelligence and later as a general tool for building "expert" and other knowledge systems. The SOAR team set about a research programme to demonstrate this, both as an effort in cognitive science and in cognitive engineering.

Despite our interest in unified theories of cognition and our sympathy for the goals of the SOAR research programme we think that much of the work that followed was conceptually and methodologically problematic. SOAR played an important part in establishing rule-based programming as a major AI paradigm (along with work on theorem provers, logic programming languages like Prolog and expert system shells), but as a contribution to the science of mind and as a platform for building artificial intelligences it is open to criticism.

SOAR was certainly an admirable attempt to capture many of the manifestations of mind with a simple computational architecture. For a while it was also a credible theory of the human cognitive system. However, this credibility was quickly eroded, particularly among psychologists who were more concerned with empirical facts about the human mind than, say, computational effectiveness or parsimony. Although Newell himself took the problem of explaining empirical observations seriously the evidence base for SOAR as a theory of the human mind was limited, and even somewhat selective, as discussed by Cooper and Shallice (1995).

In fact, from our perspective as traditional "natural scientists" the SOAR programme did not look like good science at all. Although Newell speaks of SOAR as an incremental, "Lakatosian" programme of research into unified theories he proposed no methodology for systematic development and testing of such theories. Indeed SOAR itself became something of an ideology (surely the antithesis of science) with its central tenets often justified on "sufficiency" criteria rather than empirical or theoretical grounds, and strongly defended against revision.

Following Allen Newell's death there seems to have been repositioning of SOAR, which is now less a unified theory of cognition and more a platform for engineering knowledge systems and applications. Even as an engineering tool, however, we wonder if SOAR has lost its way. As the field of knowledge engineering has moved on to develop powerful high-level tools for AI applications, like deductive databases, object-oriented knowledge bases and agents. SOAR has remained a "pure" production rule technology. Given Newell's emphasis on the importance of the "knowledge level" for cognitive science it is also surprising that successful knowledge engineering techniques have not been incorporated into SOAR. We are thinking here of techniques like generic tasks and ontologies, and formal specification and verification of knowledge bases (see Fox and Das, 2000, for a review).

In the end of course the UTC community might reasonably argue that SOAR's significance is that it shows how you can build a mind from many little parts" to quote Marvin Minsky. It embodies an "architecture for general intelligence" constructed from the most minimal elements: goal-based problem-solving, simple if-then rules, a recognise-act rule execution cycle with conflict resolution, and learning by chunking new rules as problems are solved and goals achieved. Unfortunately even this claim is ambiguous, since none of these mechanisms is demonstrably *necessary* for general intelligence. Furthermore implementations of SOAR were generally (very) large LISP or C programs, in which the basic computational principles of the theory were mired in implementation details whose role and importance for its capabilities were hard to assess (Cooper, Fox, Farringdon and Shallice, 1996).

To summarise, we are not convinced that the research methods that have been used by the SOAR community are sufficient to provide a firm scientific basis for understanding naturally occurring minds, or a rigorous engineering methodology for building artificial ones. However we do believe that the goals of the SOAR community were important and we don't wish to pick on SOAR as uniquely unsatisfactory in these respects. In fact the AI community generally has had a tendency to fall between the stools of good empirical and theoretical science and good practical engineering. However, if we are truly moving into a new period of research into minds and other kinds of unified theories then we need to adopt more systematic and articulated research methods than we have in the past.

3 Empirical programmes for understanding minds

Despite the doubts about SOAR we continue to share Newell's general aspiration to develop a unified theory of cognition, and his conviction that symbol processing is sufficient (and in some respects necessary) to understand general intelligences, and the human mind specifically. Our own work has been aimed at similar goals with rather different methods. This has emphasised systematic programmes of investigation into intelligent, knowledge-rich behaviour, seeking insights from how intelligence is manifested and constrained in complex, real-world settings.

If we can be allowed a little post *hoc* rationalisation we have adopted a programmatic approache, which can be summarized as a number of steps:

1. Define a range of cognitive functions that together are taken to embody "mind" (the theoretical *scope)*
2. Select a subset of these functions for study (a tractable *focus* for the investigation)
3. Adopt a field of knowledge (as a *domain* for the research programme)
4. Carry out objective studies of behaviour in the domain (keep the programme honest)
5. Develop a computational theory for the selected focus
6. Assess the theory against objective data and appropriate evaluation criteria

We have pursued two complementary research programmes based on this slightly sanitized model. One has been concerned with human reasoning and decision making (mostly in medicine), and one has been aimed at developing practical tools for supporting reasoning, decision-making and action (also in medicine). The motivation for this two-pronged strategy is that "mind" is a complex, multi-faceted concept that probably cannot (as yet) be understood within any one explanatory framework. Consequently a full understanding can only be achieved by exploring it with multiple theoretical tools and concepts.

Three established frameworks for understanding minds that have been developed in cognitive science are illustrated in figure 1. Our own work is attempting to address the unification of problem-solving, reasoning, decision-making, planning and other cognitive functions from epistemic, functional and structural perspectives, though we shall only discuss two programmes in this paper. These are concerned with *architectural* issues and the application of *knowledge* by different architectures.

The first research programme is addressed to the *knowledge level* seen at top of figure 1. (We prefer the term "epistemic", as this subsumes ontological and cultural as well as individual knowledge.) This has been specifically considered in the domain of medicine but the emphasis has been on building computational models of realistically complex expertise that can be put to practical use in any domain. The programme instantiates the general schema above as follows:

1. Theoretical scope – perception, memory, knowledge representation, reasoning, problem-solving, decision making, planning, action control, and learning
2. Focal subset – reasoning, decision-making, scheduling, acting
3. Knowledge domain – patient care in cancer and other complex medical domains
4. Observational studies – development and evaluation of medical decision aids
5. Theory – logical models of knowledge and expertise
6. Evaluation – effectiveness of decision aids on real world (medical) problems

The second programme follows the SOAR community and others in psychology in trying to understand the *human information processing* architecture (lower level in figure 1), focussing on how people learn to carry out reasoning and decision-making in tasks similar to medical diagnosis. The programme instantiates the general research schema as follows.

1. Theoretical scope (as before) – perception, memory, knowledge-representation, reasoning,

problem-solving, decision making, planning, action control, and learning

2. Focal subset – memory, reasoning, decision-making, action control, learning
3. Knowledge domain – medical diagnosis
4. Observational studies – controlled experiments of human subjects carrying out a simulated diagnosis task
5. Theory: computational model of human information processing architecture
6. Evaluation: comparison of behaviour of human subjects and model

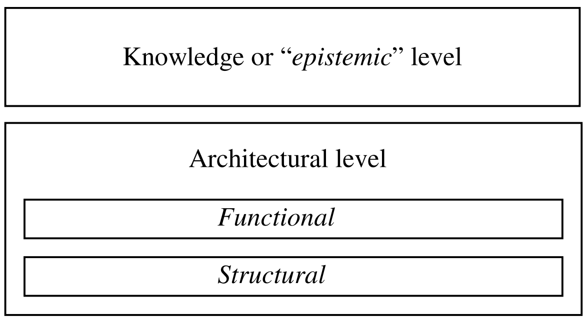

Figure 1: Explanatory levels of cognitive processes

We distinguish here between functional level of description (dealing with cognitive functions like memory, reasoning, perception and the like) and structural levels of description that are concerned with implementation in brain tissue, silicon technologies or whatever. Our own concern has been primarily with functional views, though minds can of course be investigated by neuroscientists and others who are more interested in their physical embodiments.

4 Results to date

The first programme of research is focussed on a "knowledge level" description of mind. In a sense we have developed a *competence* model of medical expertise without incorporating performance limitations such as people's limited information processing capacity or our tendency to forget things. Types of expertise have ranged from decision making in drug prescribing and risk assessment to the execution of complex clinical procedures such as breast cancer screening and chemotherapy which involve many tasks carried out over time (Fox and Thomson, 1999). As we addressed wider and more complex domains a broad model of expertise emerged, expressed as a generalised epistemic model, that is summarised in figure 2. To clarify this informal model the framework has been formalised using a combination of classical and non-classical logics to define the semantics of the arrows. These represent the kinds of inference by which goals are derived from beliefs, actions from plans etc. (Das et al, 1997; Fox and Das, 2000).

The second research programme is aimed at a functional model of the human information processing architecture (Fox and Cooper, 1997; Cooper refs). In contrast to SOAR, which makes strong commitments to a particular architectural theory, we have used a model-building environment that has been designed to permit the construct and explore the properties of alternative possible computational architectures. This system, called COGENT (Cooper and Fox, 1998) provides a set of standard components for modeling information processing architectures as a set of distinct computational compartments, such as I/O channels, memory buffers, production systems, prolog programs, and connectionist and other learning processes. COGENT provides a range of tools for programming the behaviour and setting parameters of the compartments, and managing communication between them. Such COGENT models are effectively platforms for "running" and exploring the behaviour of an expertise model under different assumptions. While the domino model in figure 2 represents a competence model of human expertise, the COGENT architecture is a performance model that can be used to superimpose features of human cognition on the disembodied epistemlc view.

Figure 3 illustrates an architecture that we have implemented using COGENT, in this case a model of how people acquire and apply knowledge in order to carry out a medical diagnosis task (Cooper et al, 1998). This model is built as a set of basic compartments that implement decision making and learning mechanisms as well as a knowledge base and working memory. Empirical studies of human behaviour on the task allowed us to identify important performance characteristics, such as the tendencies to focus on positive rather than negative information, to be overly influenced by recent information, to forget things, and so forth. The resulting model provided a successful simulation of many aspects of human performance on the diagnosis task.

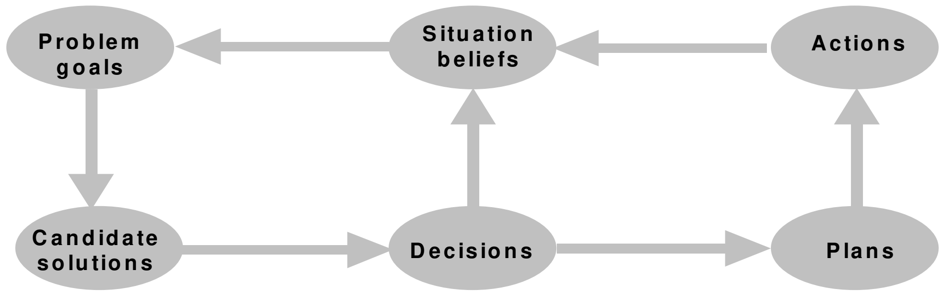

Figure 2: The "domino": a generalised model of reasoning, decision making and planning

The COGENT environment also provides explicit support for systematically exploring alternative theoretical positions. It has proved to be a highly effective tool for comparing models of reasoning and decision making in medical diagnosis, specifically comparing a family of symbolic models of medical judgement similar to the domino model with Bayesian and connectionist models (Cooper Yule and Fox, submitted).

5 Designing a mind: a software-engineering approach

In other work we have shown that symbolic models of reasoning, decision making etc. developed for modeling human information processing and as well as the decision models used in the development of practical decision aids are compatible with the broad framework embodied in the domino model (Fox and Das, 2000). In fact this model seems very general, covering as it does beliefs, goals, decisions, plans, actions and so forth, and we believe therefore that it embodies a simple, but interesting, class of cognitive agents or "minds".

To explore this, we have developed the model further, by embodying it in another development environment whose function is to support the development of expert software agents as distinct from architectures. In order to do this we decided to move from the descriptive cognitive modeling and general AI programming supported by COGENT, to adopt formal design and implementation techniques which we felt would simplify the development of agents and yield insights into their competencies and limitations at the epistemic level. The result, the PRO*forma* development environment (Fox and Das, op *cit)* can be used as a tool for conceiving, designing and implementing cognitive agents based on the domino framework, and potentially for systematically investigating the strengths and weaknesses of such agents.

The domino is primarily concerned with the *logical* processes involved in executing plans, making decisions and scheduling primitive actions. The latter include actions whose goal is to modify the agent's environment *(via* another agent, or a physical device or effector) and actions whose function is to obtain information about the environment (indirectly from another agent or directly via a sensor).

The first stage in turning the model into a practical design tool was to develop a knowledge representation language that embodies the required concepts. In common with other formal knowledge representation languages PRO*forma* is based on first-order logic (van Harmelen et al, 1993) with specific extensions to support constructs like beliefs, goals, decisions and plans. Van Harmelen and Balder (1992) summarize advantages of using such languages for describing the structure and/or behavior of knowledge systems as follows:

- the removal of ambiguity
- facilitation of communication and discussion between designers
- the ability to deduce properties of the *knowledge* independently of a specific *implementation.*

The second stage is to support this language in a development environment that provides a range of computer-aided software engineering tools that support the specification of agent systems in the language. The approach that we have taken to this is to build the environment around a basic ontology of goal-based activities, or tasks (figure 4). This ontology includes four classes of task only: *decisions* (any kind of choice); *actions* and *"enquiries"* which

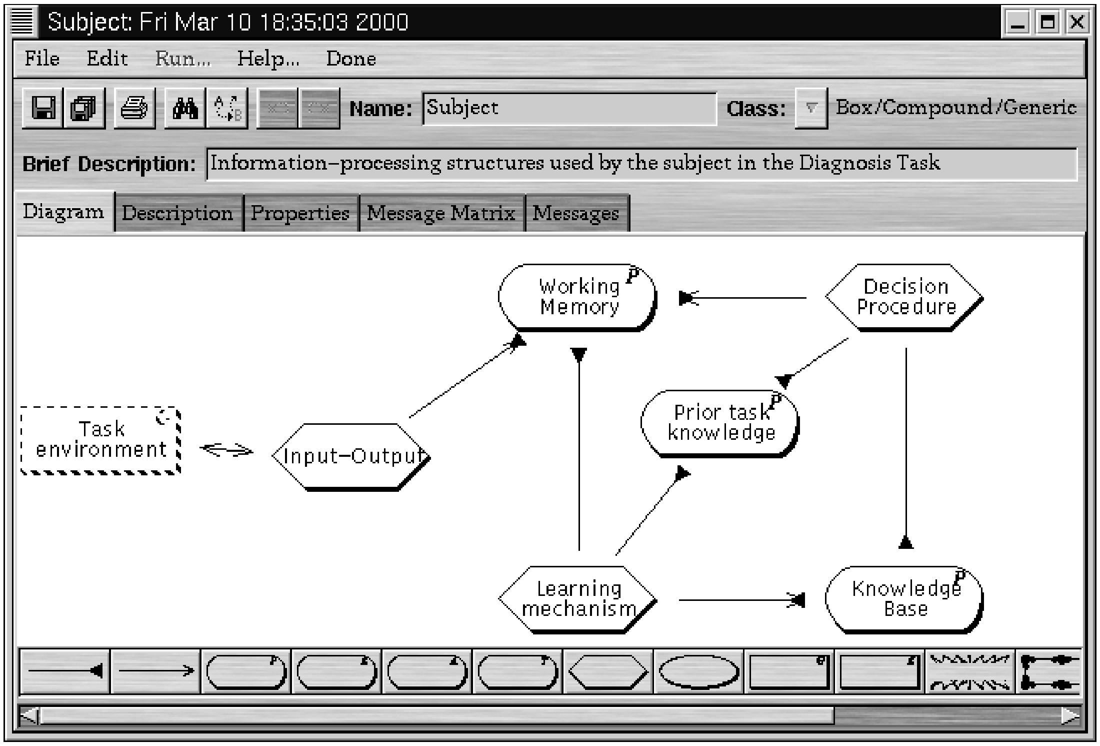

Figure 3: Information processing architecture for simulating complex cognition

are actions that are intended to modify the agent's environment in some way or return information to an agent, and *plans* (collection or sequence of tasks, including subplans). All tasks share a number of fundamental attributes, where the *goal* of the task, its *preconditions* (which must be true before the task can be executed) and *postconditions* (which will hold if the task is completed successfully) are defined. The general attributes are inherited from the generic or *root* task and particular instances of a task type (e.g. a diagnosis decision, a risk assessment decision) are defined by assigning distinctive values to the attributes. The 4 task subtypes also have distinguishing attributes. For example plans have *components* and *termination* and *abort conditions* while decisions have *candidates* and *argument schemas* for arguing the pros and cons of different candidates (Fox and Das, 2000).

The relationship between the task ontology and the domino model is shown in figure *5*. The processes on the left-hand-side of the model represent a generalised process of decision making, by which an agent commits to new beliefs and/or plans, and plans decompose into component plans and actions through a process of task scheduling.

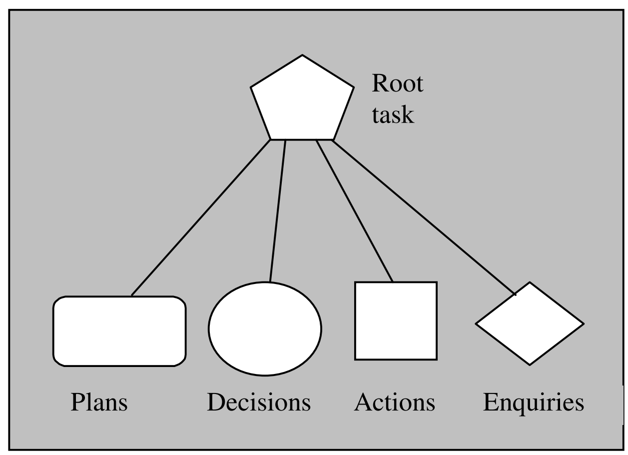

Figure 4: a task ontology for modelling expertise

The PRO*forma* knowledge representation language is a formal, compositional, specification language. It has proved to be an effective formalism for representing complex expertise models. Development of PRO*forma* agents language is supported by a development environment (developed in collaboration with *Infer*Med, Ltd). This incorporates a range of CASE tools for assembling agents from epistemic components and supports a systematic development method (figure 6).

The PRO*forma* language, toolset and development method have been validated on a range of practical applications with good results: the task ontology is small but highly expressive and the tools and method considerably simplifies the job of building epistemic models.

Some of the apparent power of the method comes from the choice of core abstractions (the four task types seem to build naturally into complex cognitive structures) and the use of proven software engineering methods (declarative specification and compositionality). Furthermore, however, the basic theoretical commitments of the language are inspired by psychology and AI (notably the core concepts of symbolic beliefs, goals, arguments, commitments and constraints). From an epistemic position these seem to capture intuitive and widely accepted ideas about important aspects of "mind".

6 Making a mind: towards a systematic research programme

Until now *PROforma* has primarily been used as an engineering tool, to build practical AI systems. However, we believe that we are now in a position to establish a new empirical and theoretical programme to systematically explore the types of agent that can be constructed within such theoretical frameworks. One such programme could draw simultaneously upon PRO*forma* as a way of modeling the epistemic level, and COGENT for the implementation level. The aim of such a programme would be a deeper understanding of the kinds of intelligences that can be implemented with this group of concepts: strengths and weaknesses, capabilities and failure modes etc.

We believe that the architectural and the epistemic programmes have both yielded significant insights into these different aspects of mind and our goal now is to integrate them. To do this we advocate a systematic programme of research in which we systematically vary competence and performance properties and parameters in order to develop a deeper insight into the underlying class of systems of which the human mind is an instance. Among the systematic variations we might carry out are the following.

Competence modelling

1. theoretical scope (e.g. extending generic task sets to include sensor and effector functions)
2. properties of tasks (e.g. symbolic, statistical or connectionist components)
3. compositionality properties (e.g. shared data versus meta-level reasoning over tasks)

Performance modelling

4. execution parameters (e.g. memory capacity, bandwidth of communication channels between compartments)
5. operational reliability (e.g. data losses during communication, memory decay)

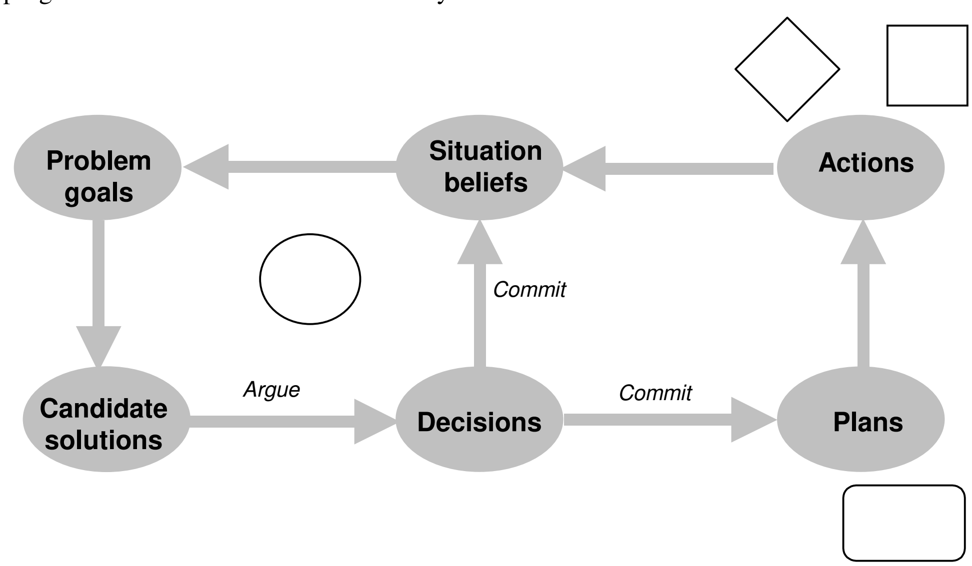

Figure 5: The relationship between the PRO*forma* task set and the functions delineated by the generalized domino model.

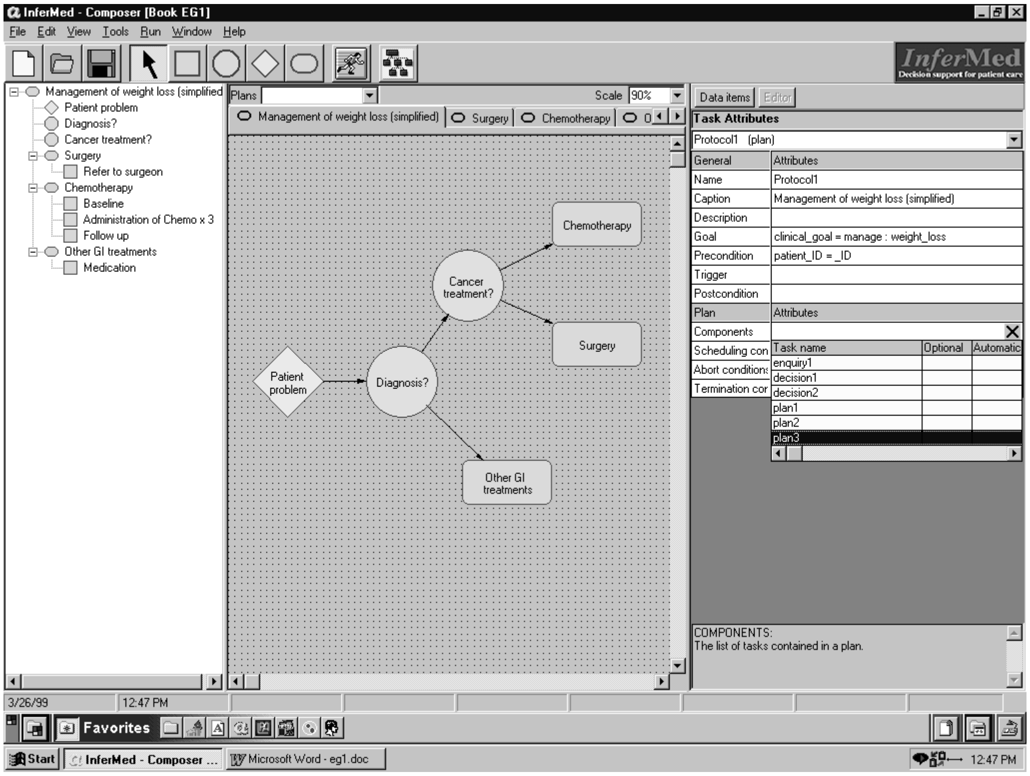

Figure 6: PRO*forma* environment for building intelligent agents based on domino model. This provides a graphical design tool for sketching a task network (left and centre panels) and populating the task definitions (right panel) using standard generic task models.

We believe that a systematic research programme along these lines would provide insights into the general nature and limitations of a significant class of natural and artificial minds.

This raises an interesting question from the point of view of the present symposium on "making a mind", viz: does *PROforma* embody a systematic method for constructing minds, albeit rather simple ones, and does COGENT offer a platform for systematically investigating the abilities and limitations of such minds? On the face of it the general approach is reasonably consistent with the broad and well-established consensus on computational aspects of mind:

Newell and Simon "We confess to a strong premonition that the actual organization of human programs closely resembles the production system organization" *Human problem solving,* 1972 p 803.

Minsky "How can intelligence emerge from non-intelligence? To answer that, we'll show that you can build a mind from many little parts, each mindless by itself." *The Society of Mind,* 1985.

Franklin "Minds tend to be embodied as collections of relatively independent modules, with relatively little communication between them". *Artificial Minds,* 1995.

Furthermore the theoretical focus of what we are doing is largely uncontroversal. As Allen Newell succinctly puts it

"Problem solving, decision-making and routine action? These are at the center of cognitive behavior. Memory, learning and skill? These equally belong at the center.... Perception and motor behavior? Matters now get interesting, because they are often excluded. Indeed the term cognition emerged in part to indicate the central processes that were ignored by peripheral perception and motor behavior" *Unified theories of cognition,* 1990, p *15.*

So the combination of PRO*forma* and COGENT seems to represent a reasonable view of what is essential to cognition, though more attention to perception and action will be needed if we are to upgrade our terminology to speak of "mind" without being accused of naievete.

Of course there will still be things missing, such as the conative and affective aspects of mind. However, in this respect we must be patient – scientific theories are always incomplete:

> "Most people still believe that no machine could ever be conscious, or feel ambition, jealousy, humor, or have any other mental life-experience. To be sure, we are still far from being able to create machines that do all the things that people do. But this only means that we need better theories of how thinking works." Marvin Minsky, *Society of Mind,* 1995.

Overall we believe that the idea that mind is composed of many little parts which act together to express "intelligence" remains a viable and arguably the most promising approach to achieving unified theories of cognition. Whether these little parts are as small as production rules or as large as, or perhaps larger than, "tasks" like those of the PRO*forma* task ontology there is much to be learned about how such entities may perform in different implementations.

7 Conclusion

For a limited but significant range of mental functions, like reasoning and decision-making, scheduling and planning, we can now give a reasonable account of how a mind like the human mind may operate, and we can build effective automata with comparable capabilities in complex, knowledge rich domains. There is much more to do, but little reason to doubt that the paradigms that are emerging can take us in productive directions.

If AI and cognitive science are to achieve their original objectives, however, we believe that they must establish more systematic empirical and theoretical programmes than have been pursued lately, and we advocate a broad agenda covering both competence and performance theories of mind. Our preferred tools are PRO*forma* and COGENT but like Allen Newell we believe a *programmatic* approach is what is important for the field. As scientists we should avoid commitments to specific concepts or tools and maintain an open mind.

References

Cooper R, Fox J, Farringdon J, Shallice T "A systematic methodology for cognitive modelling" *Artificial Intelligence,* 85, 3–44, 1996.

Cooper R, Fox J "COGENT: A visual design environment for cognitive modelling" *Behaviour Research Methods, Instrumentation and Computers,* 1998, 30 (4), 5 53–564.

Cooper R, Fox J, Learning to make decisions under uncertainty, *Proc. 19th Annual Conference of the Cognitive Science Society,* 1998.

Cooper R and Shallice T "SOAR and the case for unified theories of cognition" *Cognition, 55,* 115–149, 1995,

Cooper R, Yule P and Fox J "Learning to make decisions: a systematic comparison of three traditions. Submitted.

Das S, Fox I, Elsdon D, Hammond P "A flexible architecture for a general intelligent agent" *Journal of Experimental and Theoretical Artificial Intelligence,* 9, 407–440, 1997.

Fox and Cooper "Cognitive processing and knowledge representation in decision making under uncertainty" in R W Scholz and A Zimmer (eds) *Qualitative aspects of decision making,* Langerich: Pabst, 1997.

Fox J, and Das S *Intelligent agents: from cognitive science to cognitive engineering,* AAAI and MIT Press (in press).

Fox J, and Thomson R "Decision Support and Disease Management: A Logic Engineering Approach", IEEE Transactions on Information Technology in Biomedicine, volume 2 no. 4, December 1998, pp217–228.

Franklin, S *Artificial Minds* Cambridge: MIT Press, 1995.

Minsky M *Society of Mind,* Simon and Schuster: New York, 1985

Newell A and Simon H A *Human Problem Solving,* Englewood Cliffs, NJ: Prentice-Hall, 1972.

Newell, A *Unified Theories of Cognition,* Cambridge: Harvard University Press, 1990.

Van Harmelen F and Balder J "$(ML)^2$: a formal language for KADS models of expertise" *Knowledge Acquisition,* 4, 121–161, 1992.

Yule P, Cooper Rand Fox J "Normative and information processing accounts of medical diagnosis" in *Proc. 2Oth Annual Conference of the Cognitive Science Society,* 1998.