AI & Soc (1992) 6: 18–26

AI & SOCIETY

What Artificial Experts Can and Cannot Do

Hubert L. Dreyfus and Stuart E. Dreyfus
Department of Philosophy, University of California, Berkeley, USA

Abstract. One's model of skill determines what one expects from neural network modelling and how one proposes to go about enhancing expertise. We view skill acquisition as a progression from acting on the basis of a rough theory of a domain in terms of facts and rules to being able to respond appropriately to the current situation on the basis of neuron connections changed by the results of responses to the relevant aspects of many past situations. Viewing skill acquisition in this ways suggests how one can avoid the problem currently facing AI of how to train a network to make human-like generalizations. In training a network one must progress, as the human learner does, from rules and facts to wholistic responses. As to future work, from our perspective one should not try to enhance expertise as in traditional AI by attempting to construct improved theories of a domain, but rather by improving the learner's access to the relevant aspects of a domain so as to facilitate learning from experience.

Keywords: Artificial Intelligence; Cognition; Connectionism; Expertise; Expert systems; Learning; Neural networks; Skill

Now that the 20th century is drawing to a close, it is becoming clear that one of the great dreams of the century is also ending. Half a century ago it seemed plausible to computer pioneers such as Alan Turing and Herbert Simon that a high speed digital computer, programmed with rules and facts, could be made to exhibit intelligent behaviour. Thus the field called Artificial Intelligence (AI) was born. After fifty years of effort, however, it now seems clear, except to a few die hards, that the attempt to use rules and symbolic representations to produce general intelligence has failed.[1] Commonsense knowledge, or better, the everyday understanding that enables people to cope with entities in the physical and social world, has turned out not to be capturable in terms of rules and features. Not only has the rationalist claim that intelligence is the product of rules for manipulating a symbolic structure representing the theoretical structure of a domain failed as a

[1]Cf. Paul M. and Patricia S. Churchland, "Could a Machine Think?", *Scientific American*, January 1990, pp. 32–37.

general theory of intelligence; it cannot even be used to produce expert systems that are as good as experts.

As the commonsense knowledge problem remained unsolved for twenty years, some researchers like Terry Winograd abandoned AI; others desperately sought some new approach. This hoped for new approach has now appeared. It is neural–network modelling – the attempt to use computers to simulate an idealized model of the brain – sometimes called connectionism. As Thomas Kuhn has pointed out, in a time of paradigm shift few members of the older generation change their minds, but, as one would expect based on Kuhn's observations, a new generation of researchers has abandoned symbolic representation in droves for this new model of how to use the computer to produce intelligence. It is the potential and limitations of this new connectionist architecture as the basis for a model of intelligence that we propose to examine.

We first need an overall view of what the goal of AI research looks like in this new mode. On the older view, if AI models are to be smarter than their builders, they must express fairly accurately the designer's theory of the domain and then use superior computational abilities to draw better inferences than humans do from this theory. On the new view, the model must have computational abilities roughly comparable to those of the modeller and be the product of greater knowledge. Like conventional models, artificial neural networks require that the designer choose the features and state variables to be used to represent a situation, as well as their values in each specific situation to be considered. Since the values of these variables can be represented as activity patterns over input neurons in a variety of ways, a particular representation must also be chosen. Furthermore, if the network is to learn from experience, the output corresponding to each input situation must be specified in terms of certain features and the values of certain variables, and these as well as their representation as patterns of activity over output neurons, must be selected. If standard supervised learning is sought as knowledge, the modeller then selects a certain number of exemplary situations to be used for training the network. Each of these situations determines an input–output pairing. A learning rule is then applied that adjust various parameters of the network, such as connection strengths, until the network produces the desired outputs for the selected inputs. The initial conditions of the adjustable parameters of the artificial neural network, together with the learning rule, the cases chosen for training purposes, and the means of representing the input and output, determine the ultimate values of the parameters of the trained network.

If simulated on a computer or implemented via an analog device, this trained network will produce an output for each new input. The most striking difference between this kind of modelling and the more conventional sort is the fact that the modeller provides a history of training inputs and the network *organizes itself* by adjusting its many parameters so as to map inputs into outputs, i.e., situations into actions, without the model builder providing any rules derived from a theory of the domain.

Thanks to connectionism with its freedom from the atomistic and rationalist assumptions underlying conventional AI, there has blossomed a new interest in phenomena that the older AI, with its model of step-by-step problem solving, had

tried to ignore, notably learning and pattern recognition. Since programming computers to be intelligent no longer requires developing a theory of some skill domain, but rather consists in writing an algorithm that enables the computer to acquire skill in such a domain without having to have a theory of it, we find a new confluence of interest in neural-net models of the mind and phenomenological models of skill acquisition.

In 1986 we proposed a model of skill acquisition that cried out for connectionist implementation.[2] On our model, skill acquisition usually begins with the student learning and applying rules for manipulating context-free elements. This is the element of truth in rationalism. Thus a chess beginner must follow strict rules relating such features as centre control, material balance, etc. After one begins to understand a domain, however, one sees meaningful aspects, not context-free features. Thus a more experienced chess player sees context-dependent aspects like unbalanced pawn structure or weakness on the king's side. A further stage of proficiency is achieved when, after a great deal of experience, one is able to see a situation as having a certain significance tending towards a certain outcome, and certain aspects of the situation stand out as salient in relation to that end. Given a certain board position, for example, chess masters conclude after a few seconds of examination that the issue is to attack or defend the king-side. Finally, after even more experience – thanks to the brain configuration produced by all the past experienced situations – an expert simply sees immediately what must be done. The chess master, for example, not only quickly sees the issues in a position; the right move just pops into his head. There is no reason to suppose the beginner's features and rules, or any other features and rules, play any role in such expert performance. One can, of course, recall the rules one once used and act on them again, but then one's behaviour will be halting and clumsy just as it was when one mastered the rules as an advanced beginner.

Seen in the context of the emerging new connectionist paradigm, our five-stage model of skill acquisition and our description of intuitive expertise accounted for the failure of symbolic AI and rule-based expert systems. It also suggested that neural-network simulation *could* produce some modicum of intelligent behaviour. But, at the same time, our account lead to a pessimistic conclusion concerning the possibility of connectionist AI. Although the neural-net approach was not based on the rationalist philosophical mistake of passing over yet presupposing skill and perception, as the symbolic one was, it seemed to us, nonetheless, that neural net simulation, just because it was holistic and open to all possible ways of associating input with output patterns, would flounder on a variation of the problem of commonsense understanding that had led to the abandonment of symbolic AI.

All neural-net modellers agree that for a net to be intelligent it must be able to generalize, that is, given sufficient examples of inputs associated with one particular output, it should associate further inputs of the same type with that same output. The question arises, however: What counts as the same type? The

---

[2]For a more detailed account of the stages of skill acquisition and the implications of this account for cognitive science, cf. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, Hubert and Stuart Dreyfus, Free Press, revised paperback edition, 1988.

designer of the net has in mind a specific definition of "type" required for a reasonable generalization, and counts it a success if the net generalizes to other instances of this type. All the "continue this sequence" questions found on intelligence tests, for example, really have more than one possible answer but most human beings share a sense of what is simple and reasonable and therefore acceptable. But when the net produces an unexpected association can one say it has failed to generalize? One could equally well say that the net has all along been acting on a different definition of "type" and that that difference has just been revealed.

For an amusing and dramatic case of creative but unintelligent generalization, consider the legend of one of connectionism's first applications. In the early days of the perceptron the army decided to train an artificial neural network to recognize tanks partly hidden behind trees in the woods. They took a number of pictures of a woods without tanks, and then pictures of the same woods with tanks clearly sticking out from behind trees. They then trained a net to discriminate the two classes of pictures. The results were impressive, and the army was even more impressed when it turned out that the net could generalize its knowledge to pictures from each set that had not been used in training the net. Just to make sure that the net had indeed learned to recognize partially hidden tanks, however, the researchers took some more pictures in the same woods and showed them to the trained net. They were shocked and depressed to find that with the new pictures the net totally failed to discriminate between pictures of trees with partially concealed tanks behind them and just plain trees. The mystery was finally solved when someone noticed that the training pictures of the woods without tanks were taken on a cloudy day, whereas those with tanks were taken on a sunny day. The net had learned to recognize and generalize the difference between a woods with and without shadows! Obviously, not what stood out for the researchers as the important difference. This example illustrates the general point that a net must share size, architecture, initial connections, configuration and socialization with the human brain if it is to share our sense of appropriate generalization.

There was also a further problem. The purely associationistic pattern recognition model of learning, adopted by the connectionists, could not explain expert consensus. That is, the connectionist model of the acquisition of the ability to behave intelligently failed to account for the important fact that even though each expert has been exposed to different cases of success and failure in different sequences, experts tend to agree in their response to a given situation.

To address the above two problems we need to again ask the question: How *does* an expert cope intelligently with a domain? Only when we understand this will we have a basis for speculating about the possibilities and limitations of artificial experts produced by neural networks.

Clearly, experience improves coping performance. In considerably lower animals it is fairly certain that trial-and-error experience directly produces synaptic and related brain changes causing raw stimulae detected by the sense organs to map into better and better physical coping responses. The changes that occur during learning almost certainly cannot be even approximately described at some higher level of abstraction such as belief, goal, or mental-domain-model modification.

Matters are more complicated and controversial, however, when it comes to skilled human behaviour. On the one hand, our trained-in and imitative social comportment and the movements of skilled physical labourers such as carpenters are probably best seen as much more subtle than, but analogous to, lower animal's coping behaviour. We neither have, nor need a mental model of our social or physical vocational skill domains in order to learn through trial-and-error and instructional example to act acceptably and even skillfully when, for example, involved in carrying on a conversation or hammering in a nail. Involvement in real situations, however, does seem crucial to this effortless and usually successful coping behaviour, for if we are given a verbal description of the conversational or hammering situation what, after conscious deliberation, we say we would do is unreliable and even varies with differently worded descriptions of the same situation.

Our detached "problem-solving" comportment when we are beginners in a new and largely cognitive skill domain or when we are faced with entirely novel situations in cognitive domains in which we have already acquired skills, on the other hand, certainly seems to be at least approximately describable at the abstract level of reasoning about the situation based on a theory of the domain. What a novice *decides* he would do when a situation is described to him in terms of what he has been taught are the salient features of such a situation is usually what he really would do in a situation with these features.

Most of the vocational activities of so-called knowledge workers fall between these extremes of novice deliberation, on the one hand, and purely intuitive manual skills, on the other. It is here that modelling is most potentially rewarding but at the same time most controversial. Business persons, surgeons, teachers etc. cope fairly effortlessly and, most of the time, successfully with situations that are hardly novel, yet not identical with one previously experienced. They do so, particularly if time is short, with no awareness of detached problem solving, and even when time permits they more often deliberate about the relevance of their prior experience and the possibility of overlooked alternative perspectives or available facts than about the rules and principles underlying their skill.

At least four models of knowledge worker's coping behaviour have been proposed: The old symbolic AI view holds that unconscious problem solving, not different in principle from that consciously used in novel situations, takes place and that, while hard to elicit, a description at the abstract level of a theory of the skill domain is possible and desirable.

A second extreme position claims that an experienced and skilled knowledge worker's cognitive processing is analogous with that of skilled physical labourers. That is, it is a brain process resulting from trial-and-error and from instructional example that is triggered by involvement in real situations and that cannot be described at any domain-theory level of abstraction.

At least two distinguishable and more plausible explanations of expertise lie between these extremes. View three (the case-based approach) holds that, in the case of experts, tens of thousands of experiences (somewhat abstracted descriptions of situations and associated successful or unsuccessful coping behaviours) are stored separately in memory, and situations similar to the one currently encountered are accessed and used to determine associated behaviour.

A fourth position, the one we hold, claims that early formal learning determines how subsequent experiences will modify involved behaviour and plays an important role in the ultimate level of skill attained, yet is unrecognizable, in principle as well as practice, in the brain processes of the skilled expert. Rather, after considerable experience in a domain, further experience produces synaptic and related neurophysiological changes not describable at a higher level of abstraction, although the neurophysiological state that these experiences modify is itself the product of the abstractly describable rule-based, theory-driven detached problem-solving behaviour taught to the beginner. Acquiring expertise according to this model, consists of a gradual brain modification, undescribable at any higher level of abstraction, that nonetheless bears the imprint of early theory-based learning.

Let us examine and evaluate the implications of these four views of human skill acquisition. Because they are most common, we shall focus on situations where experienced experts face situations similar to, but not identical with, ones previously studied or experienced.

If one holds extreme view number one, that expert behaviour is the result of problem-solving based on a theory of the domain, conventional AI seems appropriate. The problem here, of course, is what domain theory to incorporate. Since experts are rarely, if ever, *conscious* of running mental models, the attempt to elicit the expert's theory of the domain turns out to be extremely frustrating. This phenomenon, among others, makes the picture of expertise explicitly or implicitly held by expert-system designers and by most AI workers exceedingly implausible, although it is not provably mistaken. Other facts calling this picture into doubt are 1) the extreme speed and ease with which experts, as opposed to problem-solving beginners, cope with their environment, 2) the failure of even the most complicated expert systems (which rely on an inference-based domain theory rather than a dynamical-system-based one) to perform at true expert level in domains where quality of performance can be verified, and 3) the established ability of artificial neural networks (and therefore presumably also of brains) to behave intelligently, admittedly currently at much less than expert level, without their computational process being interpretable as the application of a domain theory. Even Herbert Simon, considered by many the father of conventional rule-based AI has, after careful observation of the phenomenon of expertise, discarded the problem-solving model in favour of one based on tens of thousands of remembered experiences.

Suppose, at the other extreme represented by position two, that the skill and expertise of a knowledge worker is solely the product of synaptic and other brain changes produced during successful and unsuccessful experiences and that the processing of input stimulae leading to output behaviour allows no abstract level of interpretation. Might the synapses and other parameters of an artificial neural net be modified by reinforcement during successful responses and by inhibition during unsuccessful ones so as to cause the net to produce expert-level responses when given new situations that are similar to, but different from, those used during training?

Efforts in this direction are continuing and it is too early to assess their success. The problem, for this approach, however, is that an artificial neural network

involves so many parameters adjustable during learning that even the learning of correct responses to tens of thousands of cases fails to determine uniquely these parameters, and hence the responses to other inputs. Given the same training cases, details such as the initial configuration of the artificial net before learning is begun, the representation in terms of neuronal activity of the input and output, the parameter modification procedure during learning, etc. can yield final trained nets agreeing on the training cases but differing completely in their responses to new inputs. If no two nets generalize similarly, there is little reason to trust any particular one's responses to new cases. The few success stories currently circulating in the artificial neural network community involve selecting and publicizing, out of many experimental networks using different architectures and training parameter values, the one that performed best on a set of additional test cases.

Position three – the case-based approach – avoids the problem of the first two views. Suppose that, as Herbert Simon has proposed, experts remember tens of thousands of situations and associated successful responses and use these to guide behaviour in new situations. Suppose further that one can represent these cases in terms of features and the values of variables and, after representing the present situation in a like manner, prior situations can be identified that are, by some measure, similar to it. Then, if the responses of these most-similar cases can somehow be combined to produce a response (or, if the responses suggested by these most-similar cases can each be separately evaluated by some criterion), stored experience can be used to generate the kind of intelligent output that conventional AI models seek to produce.

The above supposition, however, leaves unanswered the question of what measure of similarity to use, how to combine possibly contradictory responses associated with similar cases, and what criterion to use if several possible responses are to be compared. Only if experts responded to situations in the manner described above is there reason to believe that answers exist to the above questions. Even then, since experts are unaware of using such a memory-based procedure, the modeller would have to guess at answers or else plumb the expert's unconscious. These practical roadblocks, plus this model's inability to explain the almost instantaneous and effortless behaviour of involved experts, as well as the fact that artificial neural networks and apparently also the brain do *not* store experiences separately and access them using a similarity measure when they learn from experience, suggest we should reject as misguided this explanation of expertise and this proposal for modelling it.

Of the four speculations about how knowledge workers learn to cope successfully, only one remains. While it is related to the above pure associative neural network model that had to be rejected due to its unpredictable generalization behaviour, it differs in important respects. Suppose that students of a skill domain initially approach the subject by means of a model of the domain. If a large group of future experts receive similar training, their ability and desire to identify certain important conceptual considerations, to assess these considerations, and to combine these assessments in order to produce predictions and decisions will be similar. At this point, rather than their brains being the *tabula rasa* assumed at the beginning of the associative training used in current neural

network models, there may well be significant similarities in the synaptic connections of all of these learners' brains. Later, when experience-based learning begins to modify the neural connections, and interpretation of the brain activity in terms of a domain theory thus ceases to be valid, the network connections at the neural level will certainly differ from expert to expert, but rarely by so much as to lead to totally different responses to a particular situation. Thus the arbitrary nature of generalization from learned cases to outputs for new cases might be avoided.

The above phenomenology of skill acquisition and the associated research program it entails for neural-net workers, while not a ground for optimism concerning the achievement of general artificial intelligence in the next century, does leave open the possibility – though by no means the certainty – that neural network research will produce artificial expertise in isolated domains. There is no obvious theoretical reason why neural network modelling could not be successful in domains such as chess playing, where basic human biological needs and desires play little or no role and imitative and trained-in human interpersonal social behaviour is largely irrelevant. The practical problems, however, are immense. No one currently has any idea how the brain, operating at the neuronal level, supports the sort of conceptual learning and thinking of which, as beginners, we are consciously aware. The current serial AI computer programs that might be said to stimulate the beginner's domain theory-based behaviour do so using a conventional program with no commitments concerning neural-level implementation. But if one is building an artificial neural network that is ultimately to modify its synaptic connections based on concrete experience, but initially is to instantiate conceptual understanding, one cannot avoid this issue.

While an artificial brain might not need to support what could be recognized as conceptual thinking in exactly the same way as do *our* brains, the conceptual understanding would need to be as rich and subtle as those achieved during human learning if further case-based experience is to produce expert or higher-level behaviour by means of synaptic and other modifications. Since most concepts of real-world interest do not admit of definition in terms of necessary and sufficient conditions, it is probably the case that the possibility of conceptual thinking about real-world situations emerges only out of real-world contextual experience. For example, as the learner experiences or studies many instances of weakness on the king side embedded in various contexts, each of these experiences producing activity patterns in the brain. The brain's pattern of activity when the subject consciously assesses a situation as weakness on the king side, then probably resembles what is common to its activity during each of these concrete experiences. This commonality might be called the concept "weakness on the king side" and it may well admit of no higher-level abstract description. If this is so, the neural network designer would need to model the process of the emergence of concepts and their combination to produce the behaviour of the learner. Then, and only then, should the process of synaptic changes based on massive further experience typical of current artificial neural network models be employed.

This very incomplete discussion of the issues that must be faced by neural network modellers if the fourth view of coping skill is correct is intended only to

suggest the immensely difficult nature of the task of capturing within an artificial neural network the learning mechanisms of a human being. Certain currently proposed "connectionist" models, where concepts are represented by individual artificial neurons and connection strengths are supposed to model relationships among concepts, fail even to begin to deal with the realities mentioned above and so offer little hope of acquiring expertise.

Conclusion

As any sports broadcast or financial newsletter shows, the computer can and does now provide us access to an almost unlimited number of facts gleaned from past data. Certain of these facts, in some situations, hold the long-term potential for improving performance. Initially, however, information that we never before had available can only either be ignored as we intuitively cope, based on past experiences where these facts were lacking or, worse, cause us to adopt newly invented rules and procedures to incorporate them in detached problem solving, thereby forsaking our intuitive expertise. Given the strong potential for initial regress that a glut of new information entails, the *immediate* challenge for the 21st century, if we wish to exploit the computer's remarkable data-processing power to enhance expertise, is to identify, in each skill domain, those computer-generated facts and displays having the property that our intuitive coping ability improves after sufficient experience with real situations where these computer outputs comprise part of the situation. Our *ultimate* challenge is to develop a theory and accompanying experimental techniques that enables us to produce, in any domain, facts and displays that improve our intuition. To pursue this research it is essential that our theory be in conformity with the phenomena of skill acquisition and expertise, that is, that it free itself from the currently entrenched rationalistic view that inferential reasoning based on a mental model or domain theory underlies all understanding and successful coping.

Correspondence and offprint requests to: Hubert L. Dreyfus, Department of Philosophy, University of California, Berkeley, California, 94720, USA