Totally Model-Free Learned Skillful Coping

Stuart E. Dreyfus
University of California, Berkeley

The author proposes a neural-network-based explanation of how a brain might acquire intuitive expertise. The explanation is intended merely to be suggestive and lacks many complexities found in even lower animal brains. Yet significantly, even this simplified brain model is capable of explaining the acquisition of simple skills without developing articulable rules for behavior or a model of the skill domain or an explicit identification of which observables in the environment are necessary for skillful behavior. Furthermore, no memories of prior experiences during the learning phase are explicitly stored and accessed during behavior. The explanation thus calls into doubt many conventional and intuitively reasonable assumptions concerning the learning and production of intuitive expertise.

Keywords: *model free; intuitive expertise; reinforcement learning; neural network*

In *Mind Over Machine* (Dreyfus & Dreyfus, 1986), my brother and I published the description of a five-stage process by which an adult learner of a new skill progresses gradually from novice to expert. The first three stages are primarily analytic and rule based, whereas the latter two stages become nonanalytic and what we call "intuitive." Intuitive expert behavior is characterized as situational responses based on what experience has shown to be successful. The question arises, How does the brain learn to produce intuitive responses?

In the 1986 original version of our book, we speculate that somehow, the expert remembers an experience most similar to the one currently encountered and that experience then guides his or her behavior. This premise creates a problem because experiences are always contextual in that some aspects stand out as salient and others are relegated to the background. In our view, however, salience depends on recognition of the appropriate prior experience, whereas recognition of that similar prior experience requires salience. We attempted to evade this dilemma by appealing to a hypothetical holographic mechanism, but neither we nor most readers found this entirely satisfactory. We called the process, regardless of how it was produced, "pattern recognition."

By 1988, when the paperback edition of the book was published, we had become more aware of both neuroscientific knowledge of the brain and of computer-based, artificial neural-network models. We came to believe that a neural mechanism could modify its synaptic connections on the basis of experience and that the modifications, based on the totality of successful experiences of the developing expert, could produce successful response to new situations that were similar to one or more previously experienced ones.

No storing of separate memories of prior experiences is required. In our new edition, where feasible, we changed the words *pattern recognition* to *pattern discrimination and association*. This explanation seemed more plausible, but the full implications of this synapse-based picture were yet to be worked out.

In what follows, I shall show how this brain-based view of the effortless and generally successful intuitive behavior of an expert in some domain, or more commonly some subdomain of a larger domain, appears to solve many of the problems posed by our earlier pattern-recognition explanation and to yield some surprising implications concerning intuition—implications missed by current studies of intuition that still depend on a pattern-recognition hypothesis (see Klein, 2003). The neural-network-based explanation of certain forms of intuitive expertise to follow does not imply that all aspects of socially embedded human behavior admit a similar form of explanation.

The discussion to follow, and most of the content of *Mind Over Machine*, assumes that a performer

Bulletin of Science, Technology & Society, Vol. 24, No. 3, June 2004, 182-187
DOI: 10.1177/0270467604264813

receives from the skill environment an objective evaluation of the level of success or failure that an action or sequence of actions produced. One wins or loses a chess game. One drives successfully in a situation or else has an accident or a frightening near miss. But this type of objective expertise is not the end of the story where socially embedded human behavior is concerned. Often, the quality of performance is subjective and depends on the performer's life history. People have differing assessments of a conversation as perhaps hurtful or informative. Behavior can be seen as socially responsible or foolishly selfless. When quality of performance is subjective, a complete story of intuitive behavior has to include an explanation of subjectively experienced quality. Why do we all experience the world differently and take pleasure or pain from our actions? Although the answer to this question lies in some combination of an individual's nature and nurture, a brain-based explanation is currently lacking and may be beyond what neuroscientists, computer scientists, philosophers, and other thinkers will ever be able to explicate fully. But the obvious fact that each of us experiences the quality of the results of our actions with various degrees of either positive or negative emotional involvement is sufficient for our purposes in what follows.

The world is composed of interacting agents, adapting to each other and to the context that their interactions help create. I restrict my attention in this article to the agents' trial-and-error learning of coping skills without expert instruction. One might, in certain domains, refer to these coping skills, when well learned, as intuitive expertise. Coping is what animals presumably do their whole life, how infants spend most of their time, what older children and adults do when learning new skills without benefit of instruction, and, perhaps most important, what adults do when exercising and improving skills (such as automobile driving) that they may have initially learned through instruction. The last of these situations is the focus of Stage 5 of our skill model, but this brain-based explanation applies equally to all of the above cases. All of the reflective, logical, and linguistic aspects of intelligence are beyond the scope of this discussion. It can, however, be argued that the sort of coping skill considered here is fundamental to these so-called higher forms of intelligence and that, ultimately, a sound understanding of higher intelligence must be built on an understanding of so-called animal-like, skillful coping.

A dog can learn to catch a Frisbee by means of trial and error. Likewise, a baseball outfielder, with sufficient practice, becomes skilled at catching fly balls. There are at least three difficulties in attempting to explain these phenomena. Because an encountered situation will rarely exactly match one previously experienced, and the appropriate response will rarely exactly match a prior response, how can one effortlessly and immediately respond? This phenomenon calls into doubt the pattern-recognition perspective because this view, when dealing with a situation not exactly matching an experienced prior one, would have to envision recognition of several similar prior patterns and combination of their responses with due weight on each. This would require effort, would be time consuming, and would presume use of some measure of similarity that, in turn, would have to depend on the nature of the situation. Secondly, how does the learning agent know, after an unsuccessful experience, which of a sequence of actions was inappropriate? Feedback concerning quality comes only after the entire sequence, and many of the actions may have been appropriate ones. This has been called the *blame or credit assignment problem*. Lastly, the stimuli received at the present moment may be insufficient to produce a successful action because the stream of stimuli and actions in the recent past leading up to the present may also be relevant to the action. For example, during the first moments of the flight of a baseball after it leaves the bat, a fielder's actions are dictated not only by the observed position of the ball but also by its velocity, which is not directly observable but is determined by its rate of change of position. This requires sensitivity to recent past observables. Business-intuitive expertise as a business plan unfolds might involve sensitivity to trends (rates at which observables are changing) and to actions taken or observed in the recent past. How, through experience, does the agent develop sensitivity to this past-dependent stream of observations?

Finding encouragement in the standard and often successful approach of the physicist or engineer to understanding and controlling a physical system and hoping to avoid the remembered-pattern-recognition hypothesis, a natural approach is to assume that the brain must somehow form a model of the dynamics of the situation on the basis of experience and use this model, together with a criterion for quality of performance, to select appropriate actions. The model must predict the effect of actions taken in situations. To do so, it must make explicit what constitutes a situation, with recent past stimuli as well as current ones possibly included.

This supposed need for a model is made explicit by Seth Lloyd (1995), who describes what he says Herbert Simon calls blueprints—descriptions of state—and asserts that "systems learn about their environment by attempting to control it, and modify their representation of the environment as a function of the results of those attempts at control" (p. 18). He asks, concerning robotics, the following:

> How can a robot that has been assigned a particular task, such as catching an irregular, bouncing ball, decide what information is important to gather, how can it best incorporate that information in a model of its task, and how can it learn to perform that task in real time? (p. 18)

A more recent approach in the spirit of rule-based artificial intelligence dispenses with modeling the environment as a necessary step toward action determination and sees the mind as using trial and error together with adaptation to find heuristic rules (sometimes called production or if-then rules) operating on environmental observables that produce successful coping behavior. A sophisticated variation on this second approach acknowledges the insufficiency of current observables alone as inputs into the heuristic rules and presumes that the brain learns from experience what past observables, together with current ones, must be used as input into heuristic rules to produce successful behavior.

Both of these approaches share one fundamental feature. They assume that through experience the mind comes to make sense of the skill environment in a way that could, at least in principle, be explained in a mentalistic language, referring to explicit rules and models, even if these rules and models are actually used unconsciously. By mentalistic rules and models, I have in mind the type of representations one is, or at least could be, conscious of, as opposed to neural activity, which is not something one can experience. To learn to cope successfully in a world shared with other intelligent agents with possibly antagonistic or self-serving goals, the mentalistic approach becomes even more complicated; the actions of others becomes relevant to one's own actions, and making sense requires assumptions, explaining, and perhaps predicting the actions of others. If, furthermore, an agent recognizes that the other agents may simultaneously be making sense of that agent's own actions when choosing theirs, the situation becomes almost hopeless, and mentalistic modelers of skillful behavior risk embarking on an infinite regress involving what one agent thinks that another agent thinks that the first agent thinks and so forth. Questions then arise concerning how the mind knows how and when to terminate this regress. Brian Arthur (1995, pp. 20-25) describes this situation and suggests that individuals form models or hypotheses to work with and modify their beliefs in currently held hypotheses on the basis of experience. The rules the mind supposedly learns and uses to do this become part of the skill-acquisition model, just as do the rules that might be used to identify those recent-past stimuli, in addition to the current stimuli, that should be used in determining actions.

This article proposes a view of the brain's processes that seems to cut through all of the difficulties raised by pattern recognition, as well as model building, information gathering, rule constructing, and hypothesis forming, and suggests that all of these assumed activities may be merely artifacts of conventional mentalistic views of skill acquisition.[1] I follow most brain modelers in taking the brain to be describable as a continuous-time, nonlinear, dynamical system composed of a massive set of interconnected neurons, each exhibiting, at any particular time, some level of activation. Activation refers to the frequency of electrical impulses emitted by a neuron. The rate of change in a neuron's level of activation at each moment depends on the neuron's level of activation and on the level of activation of neurons connected to it by synaptic connections. During learning, these synaptic connections are altered. Certain neurons receive input stimulation from the organs sensing the environment, whereas others are connected to effectors that produce the organism's experiences or actions. Most of the neurons act as intermediaries between input and output and are called hidden neurons. At any moment, their level of activation determines how incoming stimuli produce changes in the output. I do not assume that the neurons' pattern of activation explicitly encodes information concerning the recent past during a particular performance of a task. Rather, I assume that the current activations of all of the neurons implicitly encode this past in that their activation pattern is determined by recently past inputs and actions during a particular performance of the task and is different for each differing past.[2] The sensitivity of the hidden neurons' current activations to the preceding events is key to much of what follows.

This view of the brain easily resolves the difficulty raised by the pattern-recognition hypothesis; namely,

that no incoming pattern need exactly match a remembered, previously encountered one, so that to produce a response, some complicated combination of various remembered situations and responses must be performed. Any input into a neural network will automatically produce an output, and, generally, inputs similar to those previously experienced combined with hidden neuron activation similar to what accompanied those previously experienced inputs will produce similar outputs.[3] Exactly what that output will be is determined by the synaptic strengths produced by the totality of prior experiences, with no calculation combining various remembered, separate experiences involved.

To treat the second difficulty mentioned at the beginning of this article (that of improving performance based on a series of actions with feedback concerning quality of performance coming only at the end of the sequence), in the spirit of temporal difference reinforcement learning, I hypothesize that certain neurons become critic neurons.[4] The activity levels of these neurons can be interpreted as determining, at each moment, the brain's critic-neurons-based expected quality of overall performance of the task given that the sensed current environment and the brain's current state, which, in turn, is a function of what has happened in this particular execution of the task up to now. Although one need not be conscious of this expectation during performance of a task, the phenomenology of skillful behavior suggests that this information is indeed encoded within the brain and is, in fact, available to the conscious mind. A baseball outfielder seems to know only moments after a ball is hit whether it is catchable and with what difficulty and likelihood. A skilled chess player, shown an unfamiliar yet realistic chess position, can report almost instantaneously the probable outcome of the game if it were contested by skilled players.

Suppose now, again in the spirit of temporal difference reinforcement learning, that, at least during the skill-learning phase, the action taken at each moment depends on the activity of a set of output neurons but that what is actually done is a small, random variation on the action dictated by these neurons. This mechanism produces the exploration needed for learned improvement. Actions may not be just passive responses to stimuli but may also involve bodily motions actively modifying what environmental stimuli are received. As stimuli arrive from the skill environment and an action is taken, the brain's expectation of quality of overall performance changes. If, as a result of incoming stimuli, current neuron activation levels, and an action, the change in the critic neurons' evaluation of the expectation of the quality of overall performance is positive, that action deserves to become a greater possibility in the future should the same incoming stimuli and brain situation occur. Synaptic changes then occur that will move the brain's future action output in that situation toward the action actually taken. Simultaneously, because the action led to the expectation of an improved outcome, the expectation that is associated with the situation in which the action was taken is increased through synaptic modification.[5] If the situation is of the sort where similar situations should produce similar actions and expectations and if the net is of the sort having the normal property that similar stimuli and neuron activation patterns produce similar output, then the synaptic changes produced by the particular event just described will also produce improved actions and expectations in similar situations, should they arise. Note that this goes on continuously as the process unfolds with the passage of time (or, sometimes, discretely in time, if the process involves distinct stages). Hence, contrary to what one might expect, the synaptic modification process constituting learning need not wait until the end of a series of actions when quality of performance becomes known.

Thus, expectations and actions are sensitive not only to the sensory input but also to the current levels of activation of the hidden neurons, and these activations are, in turn, functions of all preceding actions and stimuli during a particular performance of the task. Therefore, the third difficulty—potential sensitivity to the recent past in addition to the present—is resolved. There is no need to interpret the pattern of hidden neural activity as representing anything in particular about the recent past; it is only required that different pasts, during a particular performance of the task, produce different patterns.

In summary, the above story avoids the separately remembered experiences hypothesis with its problems in dealing with situations not exactly matching any memory. Nor does it require any mental models or heuristic rules. It also renders action improvement during learning possible at each moment of the performance of the task because every action immediately receives from the critic neurons either credit or blame. Furthermore, by making the actions and expectations dependent on the past actions and stimuli through the activity of hidden neurons, as the task is performed, behavior becomes sensitive to the past as well as to the present without asking or answering any questions

about what in the past is relevant. Nor must the relevance of incoming stimuli be learned because, due to an averaging-out process of what the brain would take to be relevant in each particular experience, irrelevant stimuli (such as the color of the batter's uniform when the task is the catching of a ball) will end up, after vast experience, with zero weight (synaptic strength) on the connections of the input stimuli to the hidden neurons of the brain. Relevance is never learned in a mentalistic sense. Also, the agent need not learn how actions of other agents are chosen or how they affect the process. The agent need only observe how the process unfolds. If the nature of the environment is slowly changing as other agents adjust their behavior, appropriate actions and estimated overall quality of performance will also slowly change because of observed quality of actual performance. What I have described can aptly be called totally model-free learning.

I hope I have made it clear why I believe that many of the difficulties and complications that arise when an agent is assumed to be either a manipulator of a vast array of separately remembered situations or a disembodied, detached decision maker trying to make a model of its environment and then respond sensibly based on that model vanish when the agent is seen as an embedded, involved, adaptive entity using only environmental feedback and its own internal state to learn to respond in a model-free way.

What are the implications of my proposal for how humans and other animals learn skill through experience? The most obvious is that if brains are modified as experiences unfold, no explicit memory of each experience need be created and stored. Hence, to ask intuitive experts to recount relevant past experiences to explain their expertise as is advocated (for example in Klein, 2003, pp. 11) is futile.[6] Furthermore, as many observers of practitioners have noticed, the expert cannot give a trustworthy explanation for behavior in terms of rules, principles, or a theory of the domain. The best explanation of the source of skilled behavior may well be that experience has modified the expert's synapses so as to produce it. But there are also more surprising implications. Although experts may be unable to explicate how their brains produced the appropriate behavior, it is common to assume that at least the expert can identify what cues dictated that behavior (Klein, 2003, p. 11). Surely, the expert must learn what trends, rates of change of stimuli and, perhaps, rates of change in these rates of change matter. But we have seen that, as long as the brain is sensitive in its performance to these sorts of things, the expert need not have acquired such information. Nor can the expert reliably report which incoming stimuli are relevant or irrelevant to performance, as claimed in Klein (2003, p. 12), because the expert cannot know which synapses connecting incoming stimuli to hidden neurons have been modified by experience so as to transmit or ignore which stimuli. It may well be that any articulable explanation furnished by an intuitive expert is pure speculation and rationalization. Even assuming that the synapse-based learned responses of an expert could be assessed and possibly improved by that expert by what is sometimes called "mental simulation" (Klein, 2003, p. 16) is dubious because such simulation of what would happen if a certain action were taken would have to draw on the sort of mental model or theory of the domain characteristic of the inferior performing first three levels of my skill-acquisition model. All of this is very bad news for anyone believing that methods for enhancing intuitive expertise can be deduced from the articulations of even the most introspective of experts. Worst of all, just as excessive respect for, and dependence on, rules and principles involved in the first three stages of my skill-acquisition model threaten to inhibit the instinctive passage to the synapse-based intuitive higher stages, belief in invalid theories and prescriptions concerning how to develop and enhance intuition may well ultimately interfere with the natural synaptic development of intuitive expertise.

Notes

1. Readers familiar with current machine-learning literature will find that the ideas of this article owe much to work on temporal difference reinforcement learning, especially the actor-critic method (Sutton & Barto, 1998, pp. 151-153). For details of the solution of example small problems using computer simulations of neural networks, see my joint papers with Eiji Mizutani on model-free learning on my Web site: http://www.ieor.berkeley.edu/people/faculty/dreyfus.html.

Eminent neuroscientists have reported that in event-related fMRI brain-scan evidence something resembling temporal difference reinforcement learning is indeed implemented in the brain (Quartz & Sejnowski, 2002, pp. 103-118 and endnotes).

See also Braver et al. (2003), McClure et al. (2003) and O'Doherty et al. (2003).

2. The activation pattern of the neurons should not be confused with the pattern of synaptic connections, which are, of course, a function of the totality of past experiences in the skill domain.

3. When appropriate, based on experiences, neural networks are also capable of producing greatly differing outputs for similar inputs and neural activation patterns. For example, a net simulating a baseball outfielder can learn, when appropriate, to stop running after a long fly ball and to turn and prepare to field the ball after it bounces off the outfield wall. Also, for classification tasks such as

distinguishing one particular dog from another, various kinds of neural nets have the capacity to produce a different classification for similar inputs. It is important to realize, however, that these distinctions are the result of the synaptic connections produced by sufficient experience and in no way require the use of a formula for what constitutes similarity, as would be needed by a modeler hypothesizing that distinct remembered patterns are the basis of classification.

4. A minor change in the story to follow can accommodate the case where there is feedback about current quality as well as, or instead of, quality at the end. In the baseball example, the primary feedback is due to catching or failing to catch the ball, but less important, additional feedback perhaps comes from the difficulty in making the catch and from any sudden accelerations required near the end of the task. Fielders prefer to accelerate rapidly initially and then run at a relatively constant speed until the catch is easily made.

5. If expectation decreases because of a trial action, presumably, the reverse of this adjustment occurs. (Given the current state of brain knowledge, it is premature to propose a specific synaptic-adjustment mechanism or a specific brain architecture.)

6. Of course, a few emotion-laden extreme experiences may be explicitly remembered and recounted, but they alone cannot fully explain the behavior produced by synaptic modifications because of the totality of experience.

References

Arthur, W. B. (1995). Complexity in economic and financial markets. *Complexity*, 1.

Braver, T. S., & J. W. Brown. (2003). Specifying the neural dynamics of human reward learning. *Neuron*, *38*(2), 150-152.

Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine: The power of human intuitive expertise in the era of the computer.* New York: Free Press.

Klein, G. (2003). *Intuition at work*. Garden City, NY: Doubleday.

Lloyd, S. (1995). Learning how to control complex systems. *Bulletin of the Santa Fe Institute*, *10*(1).

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339-346.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329-337.

Quartz, S. R., & Sejnowski, T. J. (2002). *Liars, lovers, and heroes: What the NEW brain science reveals about how we become who we are*. New York: HarperCollins.

Sutton, R. S., & Barto, A. C. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Stuart E. Dreyfus, an applied mathematician, is professor emeritus in the Department of Industrial Engineering and Operations Research of the University of California, Berkeley. He coauthored the book, Mind Over Machine *with his brother Hubert L. Dreyfus and has authored or coauthored three books on dynamic programming, a mathematical optimization technique. Much of his research concerns the use, and the limitations, of mathematics and computers to aid or replace human decision making.*