#F859

In K.A. Ericsson (Ed.), The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games. Mahwah, NJ: Erlbaum Associates, 1996.

6

Perceptual and Memory Processes in the Acquisition of Expert Performance: The EPAM Model

Howard B. Richman
Fernand Gobet
James J. Staszewski
Herbert A. Simon
Carnegie Mellon University

The performance of experts in various domains has become an important topic of psychological research, especially since the pioneering study of chessmasters by de Groot (1946). This research has shown that expertise depends on acquiring large stores of relevant knowledge that are then accessible for use within the expert domain. The magnitude and nature of this knowledge has been investigated, as well as the way in which experts come to acquire it (Ericsson & Staszewski, 1989). Today we know that expert behavior combines (a) the abilities to recognize key features of situations and to access information in memory that is relevant to them with (b) the ability to solve problems by heuristic search in appropriate problem spaces (Newell & Simon, 1972).

The first ability underlies expert capacity for solving problems of familiar types rapidly and without much explicit analysis (the terms *instantly* and *intuitively* are often applied). The second ability underlies expert capacity for solving problems that require more systematic, and sometimes very extensive, analysis. In practice, most problems that an expert encounters

call on a closely interwoven combination of recognition and search processes. Because both kinds of processes draw extensively on domain knowledge, experts are usually unable to behave "expertly" when confronted with problems outside their domains.

This general, but not very precise, characterization of expertise as revealed by research explains the general mechanisms and processes that make expert behavior possible. It does not explain how the mechanisms and processes of the human brain support the expert's functioning; nor does it explain how expert knowledge and processes are learned. To link the phenomena of expertise with the mechanisms for their implementation and with the expert's knowledge base, we need more than a description of the phenomena; we need rigorous models of cognitive processes that are consistent with both what is known about expertise and our general knowledge of human perception, memory, problem solving, and learning. Such models now exist, in rather extensive form, at the level of information processes (symbolic processes). As yet, we have little knowledge about how the information-processing models are implemented neurologically.

In the first section of this chapter, we enlarge the very broad picture we have just sketched of the nature of expertise. In the second section, we describe a model of human perception, learning, memory, and search that provides an explanation for expert behavior, and of the learning processes that are used in the acquisition of expertise. The models of expert performance and learning employ information-processing mechanisms whose presence, parameters, and functioning in human cognition have been validated by a substantial body of converging evidence, most of it not derived directly from research on expertise. (For the perception and memory part of the picture see especially Feigenbaum & Simon, 1984, and Richman, Staszewski & Simon, 1995; for the problem-solving part, see Newell & Simon, 1972, and Newell, 1990).

We have little to say, except by way of occasional comment, about the realization of these cognitive processes in the physiological mechanisms of the human brain, for, as we have remarked, the linkage between the information-processing level and the level of neuronal structures and processes is still very sketchy and incomplete.

THE NATURE OF EXPERT PERFORMANCE

Definition of Expertise

In research on expertise, an expert is usually defined in a very pragmatic way as someone who performs at the level of an experienced professional: an MD in medicine, a Master or Grandmaster in chess, an experienced systems programmer, a practicing attorney, an engineer employed in design, and so on. The difference between the performance of experts defined

in this way and of novices, who lack the training and experience of the experts, is so great that it is easy to observe and characterize expert–novice differences even with this crude division; in fact, the performance of the two groups seldom overlaps.

Chess has been a very valuable domain for research on expertise because, among other reasons, there exists a standard quantitative scale for measuring differences in chess skill. The ELO rating, which is assigned to all persons throughout the world who play in tournaments, measures the player's results from competitive play against other rated players.[1] A rating of 1,800 to 1,999 assigns a player to Class A, 2,000 to 2,199 to Expert, 2,200 to 2,499 to Master, and 2,500 and over to Grandmaster rank. The ratings are so adjusted that a player whose rating exceeds another's by 200 points should defeat the latter in about two games out of three. In most research on expertise in chess, Masters and Grandmasters are regarded as experts and players of Class A and below as novices, but the research is not limited to simple expert–novice distinctions, for it can observe differences all along the scale and confirm the fact that there are not discontinuities, but a smooth gradient to the very top levels.

Some researchers on expertise (e.g., Simonton, chap. 8, this volume) would distinguish two levels of experts: those who represent state of the art practice, and those who are the discipline's creators, continually changing the discipline by contributing new knowledge, theories, and techniques. We later have some comments on this distinction between experts and creators, but work mostly with a simple distinction between experts and novices, as most of the research has been reported in those terms.

Performance of Experts

Quality of Performance

The tasks put to experts in experiments generally involve making decisions or solving problems. We combine both activities under the term *problem solving*. The consistent (and unsurprising) finding of research is that experts can solve problems in their domain that novices cannot solve, or, in the case of problems solvable by novices, experts can solve them much more rapidly and accurately. When think-aloud protocols are taken of problem solving, experts' protocols are generally briefer than novices', probably reflecting the fact that many of the experts' subprocesses have been automated so that they no longer require conscious attention. Novices tend to work backward from the problem goal, whereas on problems that are very easy for them, experts work forward, simply noticing and reporting conse-

[1]There are both international and national ELO ratings, which differ slightly, but we do not need to be concerned with these differences.

quences of the "givens" of the problem until a solution appears. On more difficult problems, experts usually revert to working backward, or at least to working in a goal-directed manner (Simon & Simon, 1978).

On problems of kinds that experts encounter frequently in professional practice, they reach rapid, sometimes "instantaneous," solutions, and are often unable to report intermediate steps in the process that led to the solution. In such cases, they commonly report that they solved the problem by intuition. In medical diagnosis, physicians often announce a tentative diagnosis immediately on initial presentation of some symptoms, but usually call for additional information before reaching a final conclusion. Similarly, in chess, experts are frequently aware of a possible move within seconds of gaining sight of a board, but will spend time—sometimes as much as a quarter of an hour—in verifying or revising the initial intuition. Similar intuitive expert behavior can be seen in other domains.

The same basic representation of a problem is frequently shared by all experts—as a product of the experts' training and experience. For example, when confronted with a problem in dynamics, expert physicists will usually try at once to express the problem in terms of differential equations. Operations research experts will classify a problem as a "linear programming problem," a "queuing problem," an "integer programming problem," and so on. Research has shown that experts generally sort and characterize problems according to the basic representations and methods relevant to solving them, whereas novices sort them in terms of surface features that often do not cue the representations that are effective for solving them (Chi, Feltovich, & Glaser, 1981).

In so-called "insight problems," the initial representation that most people adopt when presented with the problem is inappropriate. (This is what makes the problem an insight problem.) There is usually a long period during which the solver attempts to use this initial representation, followed by frustration and then attempts to find a better representation. If one is found, its discovery may be accompanied by an "aha!" (literally), and the "aha" soon followed, in turn, by a solution to the problem (Kaplan & Simon, 1990). Of course what is an insight problem to a novice may not be an insight problem to an expert, who may recognize at once from the problem statement what representation will lead to a solution. In general, subjects are unable to report the reasoning (if that is what it is) that leads them to discover the correct problem representation. The "aha" is evidence that the solution obtained suddenly was unanticipated before the change in representation was made.

Speed of Performance. Experts not only solve domain problems that novices cannot solve, but they generally solve them much more rapidly. Protocol data indicate that many subtasks that require novices to carry out a sequence of steps and to engage in heuristic search are solved by experts in a single step triggered by recognition of the appropriate cue. Thus, the

expert has replaced the sequence of operations used by the novice by some kind of macro-operation. In fact (although this may not hold for some kinds of motor tasks), superior speed in performance can generally be accounted for by "chunking" of this general kind, and does not require that the primitive processes (reaction times, times to apply simple arithmetic operators, and so on) of the expert be speedier than those of the novice.

For example, from think-aloud protocols, information can be obtained about the processes used by lightning calculators, and about the differences between those processes and the ones used by people with ordinary abilities in calculation. When allowance is made for the differences in strategy (differences that are based in large part on numerical knowledge that the lightning calculator possesses and novices do not), most of the difference in overall speed can be accounted for without postulating that the lightning calculator adds or subtracts single digits substantially more rapidly than novices do, or has unusual memory capacity (Dansereau, 1969).

Performance on Extradomain Tasks. The superior performance of experts does not carry over, in general, to tasks lying outside the domain. Of course if the extradomain tasks share common elements at an abstract level with tasks in the domain of expertise, there may be more or less extensive transfer of heuristics. (A physicist will probably have an easier time reading a text in mathematical economics than will a person without training in calculus.) But unless identifiable opportunities for transfer are present (common elements), the expert is likely to perform only a little better, or no better, than the novice.

A classical demonstration of this fact that has received much attention compares the respective abilities of chess experts and novices to replace the chess pieces on a board that they have seen for only a short interval (5 seconds, say). The boards used in these experiments typically have about 25 pieces on them, either (Condition 1) arranged as they were in a game between strong players or (Condition 2) arranged at random. In the game positions, the experts show an enormous advantage over novices: Masters and Grandmasters are usually able to replace the pieces with 90% or more accuracy (23–25 pieces, say), whereas Class A players will generally replace only 6 or 7 pieces. In the random positions, novices (Class A players) will do a little more poorly than with the game positions (3 or 4 pieces replaced correctly), and experts will average only about 1 piece more than the novices (Chase & Simon, 1973; Gobet & Simon, in press).

This experiment refutes the hypothesis that the superior performance of the experts is due to a superiority of domain-independent perception and/or memory. Instead, it appears that novices see individual pieces on the board, whereas experts see familiar constellations (chunks) of pieces—a half dozen of them, say. If short-term memory (STM) is limited by the number of chunks (familiar patterns) that can be held, then the results of this experiment can be explained without positing any difference in the

capacities of the expert and novice STMs: Both can hold six or seven chunks in memory, and when familiar chunks are rare (as on the random boards), experts do nearly as badly as novices. This same effect of chunking on STM capacity has been demonstrated for a variety of other tasks besides chess.

The Knowledge of Experts

We have already suggested that, to a major degree, the superior performance of experts can be explained in terms of their superior domain knowledge. There have been a few direct estimates of the extent of that knowledge, the estimates being all of the same order of magnitude. First, we have some measurements of native language vocabularies of children and adults. In round numbers, the measured vocabulary of a college-educated adult is usually in the range of 100,000 to 200,000 lexical items whose meanings would be understood if they were seen in the context of text. This estimate is consistent with the vocabulary size of good bilingual dictionaries, typically in the neighborhood of 60,000 words, which need to be supplemented by technical vocabularies in special domains. Estimates have also been made, by several more or less indirect routes, of the number of chunks (familiar patterns of pieces) held in memory by chess experts. These estimates lie in the range from 50,000 to 100,000.[2]

Of course, the knowledge of the expert is surely not limited to a vocabulary of familiar patterns. It also includes a more or less extensive store of representations that can be used in solving problems, actions that can be taken in the solution process, and a variety of other components. The important overall point is that the expert's ability to perform in his or her domain of expertise rests solidly on a large accumulation in long-term memory of knowledge that is (often) evoked when it is relevant for solving the problem at hand. If we broaden the definition of chunk to encompass knowledge of all of these kinds, then we can estimate that the expert holds in memory hundreds of thousands, or even a few million, of such chunks.

Becoming an Expert

The empirical studies of how someone becomes expert in a domain proclaim loudly and consistently that experts are made, not born. This is not to say that there are not innate differences among people in "talent" or ability. Of course there are. No training regimen is likely to make someone born with an IQ of 50 (using any standard measure of IQ) into a competent attorney, aircraft mechanic, flute player, or research mathematician. However, innate talent or ability only becomes expertise when it is nourished by

[2]Holding (1985) argued that the number is much smaller, but Gobet and Simon (1995), on the basis of a re-examination of the evidence, found strong evidence for the earlier estimates.

extensive training and practice. This fact has been confirmed at the highest level of expertise by studies of more than a dozen expert domains, including chess playing, musical performance, swimming, tennis, musical composition, experimental science, mathematical research, and others (Bloom, 1985; Ericsson & Charness, 1994).

World-class experts may be defined as the top few hundred persons in any domain: Olympic winners in sports, concert pianists who win international prizes, strong chess Grandmasters, Nobel Prize winners in science, members of national academies, and the like. The central research finding is that no one becomes a world-class expert without 10 years or more of intense attention to training and practice in the domain of expertise (Bloom, 1985; Hayes, 1988). Einstein, for example, was studying physics (and had even written an unpublished paper on electromagnetism) 10 years before the famous year (1905) in which he published, at age 26, his first paper on special relativity.

Moreover, even at very high levels in a domain, there is a strong correlation between acknowledged expertise and cumulated learning time (Ericsson & Charness, 1994). It has gradually become clear that neither child prodigies nor so-called "idiot savants" are exceptions to this rule: When their histories are studied, it is found that they have put in their 10 years before reaching world-class levels. Mozart, for example, was composing at age 4 or 5, but his first works that would be regarded as world class were composed when he was at least 17 (and perhaps none written before he was 21 really qualify as world class)—an interval of 12 or more years. The same is found to be the case for other prodigies.

Idiot Savants. As for idiot savants (who may form an exception to the "tested-IQ-of-50 rule" proposed already), the "idiocy" stems from the fact that almost all of their efforts have been directed toward their domain of pre-eminence, hence they know little about anything else. Their expertness is almost always acquired in domains (arithmetic is one example, music another) in which they can increase their knowledge by constant, and usually solitary, mental activity (e.g., memorizing large numbers of prime numbers and various short-cut computations that use such knowledge, or memorizing both specific pieces of music and characteristic "chunks" that lie at the basis of musical pattern). They do this without necessarily being instructed explicitly or having access to books. When information is available about the ways in which savants spend their time, there is always a history of intense preoccupation with the domain of expertise. In the rare cases where they are world class in that domain, their histories do not violate the 10-year rule.

In contrast to idiot savants, most persons who become world-class experts receive a great deal of instruction along the path, first from parents, teachers, or coaches available in the immediate environment, then from progressively more sophisticated teachers and coaches combined with

immersion in a culture that provides many opportunities for competition and collaboration with other experts, and observation of their performances.

Talent and Expertise. Sternberg (chap. 15, this volume) has argued passionately and persuasively that innate talent is essential to acquiring high levels of expertise. We must emphasize again that the empirically based 10-year rule of thumb and the 50,000-chunk estimate for the acquisition of world-class expertise set necessary conditions for these attainments, but we do not claim that they are sufficient conditions.

To establish sufficient conditions for expert performance, we have to conduct experiments, either with human subjects or computer problem-solving simulators or both, that start with rather complete information about initial memory inputs and abilities and show that these were adequate to produce some kind of expert behavior. We mention two examples where this strategy has been employed to some extent, one involving computer simulation, the other involving human subjects.

The computer program BACON (Langley, Simon, Bradshaw, & Zytkow, 1987), given exactly the same data as Kepler had available (periods of the solar planets and their distances from the sun), and no other knowledge or hypotheses about astronomy, concluded, within a short time and after generating and testing only four different hypotheses, that the data were described by the law: Period = $(\text{Distance})^{3/2}$, which is precisely Kepler's Third Law, a major landmark in the history of astronomy. We know exactly what knowledge BACON had. In addition to the data (identical with Kepler's), it had only some heuristics (rules of thumb) for generating simple functions and modifying them successively on the basis of the kind of fit or misfit it observed between function and data. Therefore, we can conclude that any system possessing these data and these simple heuristics has sufficient "talent" to make some discoveries of this kind and magnitude. Of course BACON did not only find Kepler's Third Law but, under exactly the same conditions and using exactly the same heuristics with the appropriate data set, rediscovered a whole host of important basic laws of physics and chemistry (Ohm's Law, Black's Law, conservation of mass, etc.).

To calibrate this performance against human capabilities, Kepler's data were given to 14 college students in the laboratory. The variables were simply labeled x and y, no interpretation was provided in terms of periods and distances, and no mention made that these were astronomical data. Nevertheless, 4 of the 14 students arrived at the law stated earlier within an hour. The 10 who failed to do so either (a) generated only a small set of candidate functions (exclusively linear functions), or (b) generated functions without using feedback from the previous attempts to guide construction of the next function (Qin & Simon, 1990).

As we do not know in detail what was stored in the students' memories prior to the experiment, we cannot make as confident a statement as we can

for BACON about the sufficient conditions for successful performance; but the experiment could easily be extended to discover what percentage of students having various backgrounds and histories of academic proficiency could be trained in a shorter or longer period of time to behave in the general manner of BACON and to achieve comparable success in law-discovery problems. Until experiments like these are carried out, we cannot make confident statements about the relative importance of talent and training in attaining high-level performance, or in determining the speed with which high levels can be reached.

Motivation and Cognition in Expert Performance

The histories of world-class experts show that high levels of expertise require not only learning but also the motivation (innate or acquired) that produces the necessary patience and persistence in training and practice. Even when the tasks on which the expertise is demonstrated are wholly cognitive, the attainment of this expertise can only be explained with proper attention to motivation. As this topic lies largely outside our own domain of expertise, we do not have much to say about it. In this case, silence does not imply unimportance.

A MODEL OF EXPERTISE AND ITS ACQUISITION

Having described some of the main characteristics of expertise that have been revealed by research, we wish to show (a) how the facts we have recounted can be explained in terms of a small number of information-processing structures and mechanisms, and (b) that these structures and mechanisms are not peculiar to the phenomenon of expertise but are the basic mechanisms that have been used to explain cognition generally—whether of experts or novices. Experts simply employ, at a very high level of skill in their domains of expertise, the basic information processes that other human beings employ at lower levels of skill, but the experts have access to much richer knowledge bases than are available to nonexperts. What we aim at, then, is a parsimonious account of expert behavior within the framework of a general account of cognition; in the words that Newell used to title his last book, at "unified theories of cognition" (Newell, 1990).

We need such a model for several reasons. The first is to demonstrate that our explanations are not being constructed ad hoc for each distinct phenomenon: that a modest number of basic mechanisms can account for them all. The second is to show that parameters of the model that have been estimated from one experimental paradigm (not necessarily related to expertise at all) retain the same values in the other paradigms, thereby providing the parsimony that is necessary if the theory is to be refutable or

testable. Splintered accounts of individual phenomena do not have this parsimony or this testability.

As the mechanisms that we need for our purposes have not yet been wholly encapsulated in a single unified theory, we have to settle for a little less: two interrelated information-processing models, one of them, EPAM, dealing primarily with perception, memory, and the processes of recognition that depend on them; the other, embodied either in GPS or Soar, an expert system that solves problems by searching heuristically through problem spaces. These models also incorporate learning mechanisms that acquire the expert knowledge and problem-solving processes, as well as the processes for generating and altering problem representations. We focus most of our attention on EPAM, for its role in expertise is perhaps less familiar than the role of problem-solving mechanisms.

The EPAM Model

The EPAM model of elementary perception and memory was built initially by Feigenbaum (1961) to account for the processes of knowledge acquisition and recall, and has since been expanded by Feigenbaum, Gregg, Richman, Simon, and others to model a wide variety of memory phenomena that have been studied in the laboratory (Feigenbaum & Simon, 1984; Gobet, 1993; Richman et al., 1995). In acquiring new knowledge, as EPAM becomes familiar with simple stimuli, it groups these stimuli into larger units called chunks, a phenomenon whose importance was first recognized by Miller (1956). A chunk is any unit of information that has been familiarized and has become meaningful (e.g., the words and phrases in one's natural language vocabulary, the objects that can be recognized when seen and named, etc.). The familiarity of a chunk relates to the ability to recognize it; meaningfulness to the information that is stored in association with it in long-term memory (LTM).

Chunking is recursive, so that chunks at any level can be grouped again into chunks at the next level above (visual features into letters, letters into words, words into familiar phrases, etc.). By testing EPAM in a wide range of experimental tasks, estimates have been obtained, and confirmed by converging evidence, for the latencies of the basic processes required to recognize a familiar item (about 0.5 s), learn the recognition tests for a new chunk (about 8 s), add a new piece of information to a LTM chunk (about 1-2 s) and retrieve stimuli (about 200 ms–2 s).

Short-Term Memory. EPAM models both STM and LTM. The most important STM for our purposes is the articulatory loop (Baddeley, 1986), whose capacity has been shown to equal the number of chunks that can be rehearsed in about 2 seconds, where the time required to rehearse each chunk is about 300 ms for the first syllable and 80 ms for each additional syllable (Zhang & Simon, 1985). What appears to be held in STM is a pointer to each

chunk, so that the contents of the memory can be rehearsed by accessing the articulatory image of each chunk via the pointer, thereby recovering its syllables. Information stored in the articulatory loop at any given time can be attended to, and this information is subject to learning—that is, gradual transfer to LTM. The time required for transfer is about 8 seconds for each new chunk, as EPAM's discrimination net is elaborated to distinguish it from chunks already familiar.

From the standpoint of expert memory, the principal significance of the parameters of STM lies in the constraint that they place on the time required to acquire new recognizable chunks. At 8 seconds per chunk, the 50,000 chunks we have estimated for the chessmaster's memory store could be acquired in a little more than 100 hours, smaller, by two orders of magnitude, than the 10 years required to reach world-class skill. How do we account for the remaining time? First, the 8-second parameter for storing a new chunk in LTM is derived from standard verbal learning experiments, where ability to recall memorized items is tested only a few minutes after they have been stored. It is well known that information stored under these conditions, without redundancy, decays very rapidly, so that only a small fraction is available after 24 hours has elapsed. The relearning and overlearning required for long-term retention can easily account for one order of magnitude more learning time, or a total of 1,000 hours.

Second, the 50,000 chunks estimated for the chess expert includes only patterns of pieces that will be recognized when they occur in a game position (including templates, which we discuss later, of typical positions that appear frequently in opening play). The expert must retain many other kinds of knowledge as well—in particular, knowledge of moves that may be appropriate when a particular pattern appears in a game, strategies for look-ahead analysis of moves, and so on. Again, it is not unreasonable to allow another factor of 10 in learning time to account for all of this additional information that must be stored in LTM. Finally, not all of the time spent in study is available for learning new chunks; in fact, as the learner becomes more and more expert, novel information that can be added to the knowledge store is encountered less and less frequently. We conclude that the parameters of learning speed that have been estimated from verbal learning studies are consistent with the estimate of 20,000 learning hours available during a 10-year period of work. (We conservatively estimate a 40-hour week!)

Visual Short-Term Memory. Most of the measurements on rates of learning have dealt with auditory material and the articulatory loop. The parameters of visual STM (the "mind's eye") are less well known, but there is no reason to suppose that the time required to transfer a familiar chunk of information from the mind's eye to LTM is very different from the time required to transfer an auditory chunk. Hence, our earlier comments about the time required to acquire the recognition capabilities that the expert

possesses can be applied to both auditory and visual chunks. For purposes of studying expertise and its acquisition, little specification is needed of the details of STM, auditory or visual, beyond the phenomenon of chunking and the time required to transfer new chunks to LTM.

Long-Term Memory. EPAM's LTM can be thought of as an indexed encyclopedia. On presentation of a visual or auditory stimulus, perceptual processes (not represented in the present EPAM or in most other current models of perception) extract a list of features from the stimulus and these features are then sorted down through a discrimination net to a terminal or "leaf" node. The different leaf nodes of the EPAM net represent the different stimuli that EPAM is able, at any given time, to discriminate. At each leaf node is stored an "image" of the stimulus, containing partial information about it (including the information that was used to sort it) and containing, also, associational links to other information about it that is held in LTM (semantic LTM). The EPAM net, then, is the index to the semantic LTM, providing access, through the associational links, to the information held in the latter. The net is not a simple tree but a network, for many paths may lead to the same leaf node, providing the redundancy that is required to recognize objects when they are observed from different angles and under varying circumstances.

Learning: Growth of the EPAM Net. When a stimulus is sorted to a leaf node of EPAM, its features may be compared with the features of the image stored there. If there are discrepancies between the two sets of features, EPAM can construct a new test node that tests for the nonidentical feature, and a new leaf node to accommodate the new stimulus, retaining the old leaf for the original stimulus (Simon, 1976). By this means, the EPAM net and the corresponding set of leaf nodes grow continually as they learn new patterns, performing the same function that is performed by the "hidden layers" of connectionist learning systems. Each new pattern that is discriminated gains its own leaf node, identifying a new chunk.

Large discrimination nets have been acquired in this manner. EPAM, in simulating a human subject who learned to repeat back sequences of 100 digits that were read to him at the rate of one digit per second, gradually "grew" a discrimination net with more than 3,000 leaf nodes (Richman et al., 1995). CHREST, an EPAM-like system for simulating the acquisition of chess expertise has acquired nets as large as 70,000 leaf nodes on being exposed to a large number of chess positions with the goal of retaining the patterns contained in them (Gobet, 1993; Gobet & Simon, 1995).

Acquiring Templates and Retrieval Structures. It has been found that human chess experts store in memory templates of many thousands of opening positions—the positions at which the game typically arrives after sequences of 10 or 20 "standard" opening moves. The templates of these

positions generally contain information about the locations of 10 or more pieces, and the presumptive locations of perhaps another half dozen. When an expert is given a brief view of a game position, he or she can then usually recognize it as corresponding to one of these templates, and thereby retain information about the locations of nearly half the pieces on the board. Moreover, information about additional pieces can be added rapidly by noticing whether they do or do not conform to their presumptive locations. We call these presumptive locations the template's *slots*, or variable places. It appears from experiments in which experts replace pieces on a briefly exposed board that information about the location of a piece can be stored in a slot in a fraction of a second.

An important technique of mnemonists, known and taught from Greek and Roman times, is to learn deliberately a *retrieval structure* that contains a great many slots in a determinate relation with each other. Once the retrieval structure is in place, each of its slots, like the slots of the chess expert's templates, can be filled with a chunk of information in a second or less. Classically, the retrieval structure took the form of a "memory palace." The mnemonist memorized the room plan of a palace, with various pieces of furniture permanently stationed in each room. Then, to store a particular list of chunks rapidly, the mnemonist simply assigned each item, in order, to one of these locations. Subsequently, they were recalled by accessing one room after another and noticing (in the mind's eye) what items had been placed there. The retrieval structure resided permanently in LTM, and could be used repeatedly to store new lists in its slots.

The expert, mentioned earlier, who recalled long lists of digits used a retrieval structure of a slightly more abstract kind (Chase & Ericsson, 1981; Richman et al., 1995). He simply imagined a hierarchy of slots, grouped by threes and fours, and assigned the digits to the successive leaf nodes at the bottom of this hierarchy. He also associated groups of digits (again groups of three or four) with long-distance running times he had previously stored in semantic memory as a result of his extensive experience and knowledge of running. As each three or four digits were read to him for the recall task, he assigned them to successive retrieval structure slots, and at the same time associated them with a familiar running time. Although the process was subject to some forgetting (it is estimated that he almost immediately forgot about 25% of the items stored) the redundancy of storage of information in more than one place allowed him to recover most of the information, and thereby to recall long lists without error.

The critical parameters in this kind of performance are (a) the ability, over some period of time, to build templates or retrieval structures in LTM, at a rate of perhaps 8 seconds for each new chunk, and (b) the ability to fill the slots in these structures with information (to be retained for a matter of hours) at a rate of less than 1 second per chunk. EPAM and CHREST, which embody such parameters, simulate closely the performance of chess and digit-memory experts in recall tasks, thus reconciling these "exceptional"

performances with what has been learned about human memories in the standard laboratory tasks on verbal learning. The effectiveness of the chunking/retrieval structure mechanisms embodied in EPAM to explain expert memory has thus been demonstrated in two quite different environments, in one of which the acquisition of retrieval structures was a deliberate mnemonic strategy, in the other, a by-product of practice and the study of chess.

Recognition and Expertise. The indexed knowledge base that is EPAM provides the mechanisms that we require to explain the commonly observed ability, discussed in a previous section of this chapter, of experts who solve many problems and deal with many situations "intuitively." The basis of EPAM's intuitions is simply its ability to recognize familiar cues (by sorting stimulus features through its discrimination net), thereby gaining access to the relevant knowledge associated with them. The physician recognizes symptoms and "intuitively" concludes (subject, perhaps to additional tests) that the patient is suffering from a particular disease; the attorney recognizes features of a contract and "intuitively" concludes that it subjects her client to certain risks. It is not a question of whether these intuitions are always reliable; they are often reliable enough to permit the expert to proceed with the task vastly more rapidly and reliably than the novice, who must employ much more tedious step-by-step search processes. The information recovered by such recognitions of cues can be very extensive; the physician may access not only the name of the disease but a large store of information about prognosis, methods of treatment, and so on. We no longer need to regard intuition as an unexplained phenomenon. It is synonymous with the familiar process of recognition, and we can simulate its workings with EPAM and other models of memory.

Insight and Learning Representations. Closely related to intuition, but perhaps requiring a slightly more elaborate mechanism for explanation, is the phenomenon of *insight*. Like intuition, insight involves sudden solution of a problem, usually accompanied by inability of the solver to explain how the solution was found, and often punctuated by a figurative or literal "aha!" What distinguishes occasions of insight from the much more common occurrences of intuition is that insight is often preceded by a shorter or longer period during which no problem solution is found, and often no plausible steps for moving toward a solution. When the insight finally occurs, it usually can be seen that the representation of the problem had to be changed in order to find the solution, but that the solution was found easily (perhaps even became "obvious") as soon as the new representation was available (Kaplan & Simon, 1990).

The key to the insight is the process of discovering a new problem representation. This new representation can already be present in the memory of the problem solver, or it can be unfamiliar. In the former case,

the reason for the delayed solution is that no clue is recognized in the situation that associates with the effective representation, so that the latter is simply not evoked, and hence is unavailable. In these situations, it is usually easy to produce the insight by providing the problem solver with a clue (often a very simple one) that produces the recognition, and thereby leads quickly to the change in representation. With experience, EPAM adds a new path from cues characterizing a class of situations to knowledge of one or more representations effective for dealing with it.

Where a representation is required that is not already familiar, insight requires problem-solving activity as well as recognition. The new representation must be discovered, then added to LTM along with the cues for its recognition. Lack of an effective representation was the principal reason for the failure of Newton's contemporaries, who were unfamiliar with the calculus, to demonstrate the law of universal gravitation. Although several of them (e.g., Hooke, Wren) conjectured a gravitational force that varied inversely with the square of the distance, none was able to demonstrate that such a force could account for the planetary motions. Once the calculus representation was available, the derivation was quite straightforward—today, students in first-year physics carry it out routinely. We have a little more to say about insight and invention in the next section.

The Problem-Solving Model

EPAM's indexed memory, the body of knowledge stored in it, and its processes for acquiring new knowledge by learning can account for a considerable part of the expert's superior abilities in memory retrieval and problem solving in the domain of expertise (and the absence of that superiority outside the domain). However, to complete the story, we must look at problem solving that requires more than recognition processes—situations in which more or less extensive heuristic search is also required. Here again, existing models of problem-solving processes like GPS and Soar provide most of the answers we need; but to see this we have to discuss the search process in a little more detail (Newell, 1990; Newell & Simon, 1972).

Solving a simple problem, say the Tower of Hanoi, is usually described as a search through the space of possible ("legal") arrangements of the disks on the pegs, from the starting arrangement to the arrangement specified as the goal. Successive states in the problem space are reached by moves, each of which, in this puzzle, amounts to changing the location of a single disk. The search is almost never random, but is guided by various heuristic rules that seek to guide the selection of the proper moves. The heuristic rules may be more or less complete, more or less correct.

When we come to more complex tasks, however, even tasks like finding the concept an experimenter has in mind to distinguish one set of objects from another, the search becomes more complex, usually involving inter-

action between two or more distinct problem spaces: the space of "instances" and the space of "hypotheses" (Simon & Lea, 1974). Suppose, for example, that a subject is presented with a succession of objects and asked to designate each one as belonging or not belonging to the concept by which the experimenter classifies them. The succession of objects constitutes the instance space, the hypotheses that the subject generates as possible classifiers constitute the hypothesis space. When the subject is told that he or she has made an incorrect judgment, the current hypothesis is usually rejected and a move made to another point in the hypothesis space. Again, this move may be guided by heuristics that are based on information collected from the previous choices and reinforcements. It has been shown that the behavior of subjects in concept attainment experiments can be explained in terms of search in the dual instance and hypothesis spaces.

Finding a scientific law that describes data obtained in a similarly cumulative fashion has been modeled in the same way by computer programs like BACON and others, using a dual search in the space of possible laws and the space of possible data observations. But why limit the process to two spaces? A scientist may search in a space of instruments, a space of experiments, a space of possible descriptive laws, a space of explanatory mechanisms, a space of problem representations, a space of research problems, and perhaps others (Langley et al., 1987).

Krebs, for example, in his search for the mechanism of urea synthesis in living organisms, selected that problem as one already recognized as important but unsolved, selected instruments and experimental procedures that he had acquired as a postdoctoral student in the laboratory of Otto Warburg, searched a space of substances as possible sources for the nitrogen in the urea, and when he discovered that ornithine and ammonia were implicated in the process, searched a space of chemical reactions to find a reaction path from inputs of these substances to the output of urea (Holmes, 1980).

Knowledge of this multitude of spaces as well as some knowledge of their structure is required for an expert approach to the research problem. As with every problem, progress will depend on processes of recognition (of possible representations, instruments, experiments, theoretical hypotheses, etc.), combined with search processes in the several spaces whenever recognition does not give an answer to the current question and a new one must be synthesized. The programs that have been built in recent years to accomplish these recognitions and searches are, again, recognizable kinfolk of programs like EPAM and the General Problem Solver (or Soar)—EPAM to make use of the knowledge base to achieve recognition of relevant information, GPS or Soar to conduct the searches in the many problem spaces. As an integral part of these processes, new learning goes on: New chunks are constructed, stored in LTM, and indexed by the cues that permit them to be recognized.

Acquisition of Problem-Solving Skill

It is clear from this account that the expert needs to acquire much more than knowledge in declarative form. Expert skill is heavily dependent on efficient processes, including strategies, planning processes, and representation-generating processes. If the processes in the expert system take the form of productions (i.e., If *<cue>*, then *<action>*), as they do in such problem-solving systems as GPS (Newell & Simon, 1972), Soar (Newell, 1990), and Act* (Anderson, 1983), then the learning mechanisms must create new productions that can be added to LTM and evoked, via the discrimination net, when the appropriate cues are present. Soar, for example, accomplishes this by storing information about the paths it has followed during problem solving so that it can recover these paths without search when the same or a similar problem later presents itself.

It is not yet clear whether learning schemes of this kind are sufficient to account fully for expert skill acquisition or whether additional mechanisms are required, but the power of learning by elaboration of a discrimination net combined with storage of knowledge schemas and new productions has been demonstrated in numerous contexts.

Expertise and Creativity

Simonton (chap. 8, this volume) hypothesizes that there may be a special group of experts in each domain who have not only exceptional knowledge and skill but also unusual capacities for inventing and adding new representations and other knowledge to the domain. These especially creative individuals may or may not be the very best performers. In chess, for example, Reti and Niemzovich were great and influential innovators who, although they were strong Grandmasters, never reached the very top of the ladder in chess competition. On the other hand, Morphy, Steinitz, Alekhine, and Botvinnik, each a world champion, also introduced important innovative ideas, whereas few important new ideas appear to have been contributed by Lasker, Capablanca, or Euwe, all world champions.

In this domain, as in others, we observe that certain experts do play a more innovative role than others without being the most highly regarded performers. In fact (one thinks of a Kandinsky in painting, a Schoenberg in musical composition) an expert may take innovating as his or her special role, and in domains like science, innovation (contributing new knowledge) actually defines the expert's central professional task. Moreover, we would, in fact, expect more than an average number of innovators among the top ranks of experts for, once the existing state of the art has been mastered, employing new ideas and practices is the principal remaining route to pre-eminence. The chess master has no choice but to display his innovations in his games (but they may be misunderstood by this contemporaries), whereas in other domains innovators may gain more permanent advantage through patenting or secrecy.

Do we require special mechanisms to account for these innovators? It appears that we do not, for as we have seen in the previous section of this chapter, innovating (e.g., finding new representations and new strategies) is itself a problem-solving task to which the tools of recognition and heuristic search in the spaces of possible representations, possible strategies, and so on, can be applied. The innovator, from this standpoint, is simply a problem solver (and "recognizer") who applies his or her efforts to changing the problem spaces in which recognition and search are carried on. The "great" innovators are those who solve the problem of bringing about the largest changes: Newton's calculus, Planck's quantum, Harvey's blood circulation, and so on.

If the tasks of innovation are problem-solving tasks, then we should be able to model them using the same mechanisms that have been used to model problem solving generally, and we have seen that this has in fact been done in such discovery systems as AM, BACON, EURISKO, LIVE, MECHEM, and the many others that can now be found in the psychological and artificial intelligence literatures (see, e.g., Langley et al., 1987). Nor are the examples limited to science. Hiller and Isaacson (1959) produced a very early program that composed original (and sometimes musically interesting) music, and the painter, Harold Cohen, has produced the Aaron program, which makes original drawings (both nonrepresentational and representational) that are aesthetically sophisticated and pleasing (McCorduck, 1991).

Of course, we can consider these programs simply as models of performance: They solve problems. But we must also take into account the extent to which discovery models incorporate learning processes and modify themselves in ways I have described earlier. Perhaps the simplest form of self-modification is EPAM's elaboration of its discrimination net, so that a program can progressively recognize a wider and wider range of stimuli. A related procedure, incorporated, for example, in theorem-proving systems, is the ability to store problem solutions and to use them in solving subsequent problems. Beyond even this are systems that can modify their own representations, processes, and strategies. In this category, I have mentioned Soar, which stores and uses strategies that have been successful in solving previous problems (Newell, 1990); and I would also call attention to the UNDERSTAND system, capable of constructing problem representations from verbal problem instructions (Hayes & Simon, 1974).

In the light of the experience we have already had with systems of these kinds—incorporating the basic capabilities of recognizing and of solving problems by search through multiple problem spaces and of learning—it is not unreasonable to hypothesize that creativity is "simply" unusually competent or admirable problem solving that accomplishes its tasks by the use of these very mechanisms. We sometimes produce work that is creative if we explore the space of problem representations, of instruments, or strategies, and so on, not limiting ourselves to the problem spaces that the current state of the art presents to us.

CONCLUSION

Research in information processing psychology began in the middle 1950s, with the construction of theories, in the form of computer simulations, of human performance on relatively simple and well-structured tasks, most of them tasks that were already familiar from the psychological laboratory. Two of these early theories, both in operation before 1960, were EPAM, a model of elementary perceptual and memory processes that was tested against the data from verbal learning experiments, and GPS, a model of problem-solving processes (emphasizing means–ends analysis) that was tested against the data from various puzzlelike tasks. As the mechanisms that were needed to explain these phenomena gradually became clearer, it began to appear that EPAM and GPS could go a long way toward accounting for the behavior of experts and explaining why experts are so much more competent than novices on tasks belonging to the expert domain. In particular, it became clear that expertise depends heavily on the possession of, and access to, large bodies of domain knowledge. Having in hand a number of systems capable of expert performance, it proved possible to devise learning schemes that showed how the expertness could be acquired.

The early models were improved and extended, and new models (e.g., ACT*, Soar, BACON) emerged in an effort to give a better account of the phenomena, but these and other systems that describe and explain how experts behave expertly are recognizable descendants of EPAM and GPS. The knowledge-accumulating and knowledge-accessing mechanisms of EPAM-like systems explain the expert's use of "intuition" in the form of recognition processes. The search mechanisms, extended to multiple search spaces and making extensive use of means–ends analysis and other heuristics, explain the expert's capabilities for solving scientific problems, including the subsidiary problems of finding appropriate representations, instruments, experiments, data, and hypotheses.

Although we do wish to convey a picture of continuity and cumulation in this chronicle of research, for we think these are fully documented by the vast and continually growing body of phenomena that have been successfully described and explained, we do not wish to convey a picture of completeness or even near completeness (no one would be fooled if we did). Lots of things have been demonstrated "in principle" that still need to be demonstrated in detail. A great many tasks (e.g., tasks in realms like language behavior and language learning, representation change, uses of visual imagery—an endless list) have only been touched. There are new challenges posed by the discovery of alternative mechanisms—connectionist systems and neural networks currently the most prominent among them.

In short, we are in a fast-flowing stream of normal science, which is rapidly gaining a broader and deeper understanding of human thinking in general and the thinking of experts in particular; and we may even expect

this stream to bring us closer to a solid linkage with neuropsychology, which up to the present has been a rather separate, and even distant world. As symbolic processes must be implemented by neuronal structures, getting a better understanding of that linkage is one of the important tasks before us—although, we hasten to add, not the only one.

The symbolic level itself continues to present us with innumerable research opportunities. It will be most interesting, as we watch these developments in the future to see how far the basic processes of recognition of familiar chunks and heuristic search through multiple problem spaces, supported by the learning processes that create the underlying stored knowledge and skill, will continue to stand at the core of our understanding of the competencies of experts.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Baddeley, A. (1986). *Working memory*. Oxford, UK: Oxford University Press.

Bloom, B. S. (1985). *Developing talent in young people*. New York: Random House.

Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141–190). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chase, W. G. & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–125.

Dansereau, D. (1969). *An information processing model of mental multiplication*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

de Groot, A. D. (1946). *Het denken van den schaker* [Thought and choice in chess]. Amsterdam: North-Holland.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725–747.

Ericsson, K. A., & Staszewski, J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing* (pp. 235–268). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feigenbaum, E. A. (1961). The simulation of verbal learning behaviors. *Proceedings of the 1961 Western Joint Computer Conference*, 121–132.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305–336.

Gobet, F. (1993). A computer model of chess memory. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* (pp. 463–468). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gobet, F., & Simon, H. A. (in press). Recall of random and distorted positions: Implications for the theory of expertise. *Memory & Cognition*

Gobet, F., & Simon, H. A. (1995). *Role of presentation time in recall of game and random positions* (Tech. Rep. C.I.P. 524). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

Hayes, J. R. (1988). *The complete problem solver* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hayes, J. R., & Simon, H. A. (1974). Understanding written problem instructions. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 167–200). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hiller, L. A., & Isaacson, L. M. (1959). *Experimental music*. New York: McGraw-Hill.

Holding, D. H. (1985). *The psychology of chess skill*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holmes, F. L. (1980). Hans Krebs and the discovery of the ornithine cycle. *Federation Proceedings*, 39(2), 216–225.
Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22(3), 374–419.
Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.
McCorduck, P. (1991). *Aaron's code*. New York: W. H. Freeman.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Qin, Y., & Simon, H. A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science*, 14, 281–312.
Richman, H. B., Staszewski, J. J., & Simon, H. A. (1995). Simulation of expert memory using EPAM IV. *Psychological Review*, 102, 305–330.
Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325–348). Hillsdale, NJ: Lawrence Erlbaum Associates.
Simon, H. A. (1976). The information storage system called "human memory." In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning and memory* (pp. 79–96). Cambridge, MA: MIT Press.
Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 105–127). Hillsdale, NJ: Lawrence Erlbaum Associates.
Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and the acoustical loop hypothesis. *Memory and Cognition*, 13, 202–207.