The Mind's Theorist

Richard C. Atkinson

Having once been a professor of psychological and cognitive sciences, occasionally I am asked to give a public lecture on the nature of the human mind. At one level, I am at a loss for what to say. Much progress has been made in understanding various aspects of the mind. There are well-developed psychological theories of perceptual and sensory systems, of the structure and control processes governing short and long term memory, of the motivational and emotional systems that affect learning, and of higher level cognitive skills involved in thinking and problem-solving. However, attempts to explain these theories soon bog down in technical details and leave one with the feeling that the mind is so complex and dynamic that a general understanding still remains beyond the reach of science.

Nevertheless, when one examines these various theories, an overriding concept emerges. In my judgment, it is key to the purpose of this essay. The concept, simply stated, proposes that the mind's highest-order function is as a *theorist*. Over time, the memory system accumulates information about the outside world; the mind's theorist's job is to sort through the information and develop a theory about that world. The mind's theorist interprets prior experiences and uses that analysis to formulate a theory that generates predictions of likely outcomes and opportunities when confronting future situations.

I use the term theory in the same sense that it is used in science. Kepler formulated astronomical laws of planetary motion. Galileo proposed the law of falling bodies. Newton's brilliant insight was to explain these two phenomena with his theory of *universal gravitation*. But Newton's theory, like all theories, was *provisional*. As physics progressed, the theory proved inadequate to predict the motion of subatomic particles or the flight of satellites in space.

Accordingly, Einstein's theory of *general relativity* replaced Newtonian theory. All theory is provisional. As more data accumulates and a wider range of phenomena are analyzed, theory invariably must be modified and updated.

As is the case for scientific theories, the mind's theory of the world is provisional. As information accumulates in memory, the mind's theorist (henceforth MT) must update the theory to account for new observations. For familiar situations, the theory is well developed and the individual responds quickly. For novel situations, a more extensive search of memory is necessary and even then may fail to generate relevant predictions. In these cases, the individual will hesitate to respond and, at the end of the episode, store additional information about the outcome.

When I was a graduate student some 60 years ago, the field of psychology was dominated by an S-R version of behaviorism. If a stimulus and response occurred in temporal contiguity and was followed by a reward (reinforcement), then supposedly a bond formed between the stimulus and response. Whenever the stimulus occurred, the bond would tend to elicit the response and become ever stronger with repeated rewards. The behaviorists were successful in explaining some simple forms of behavior, particularly that of laboratory animals. They argued that even the most complex forms of human behavior could be explained by postulating a hierarchy of stimulus-response associations, once the basic associations were better understood.

Pavlov's classical conditioning experiment was an example of an S-R approach to psychological phenomena. A puff of meat powder (delivered to a pan in front of a dog) would elicit salivation. Next, a brief tone was paired with the presentation of the meat powder. After a series of such trials, the tone (when presented alone) led to salivation. Experiments on Pavlovian conditioning proved to be very productive and placed psychology on a firm base as an experimental science. However, by the 1960s it became evident that an S-R analysis could not

explain many of the newer research findings. A cognitive interpretation was necessary. For example, in the Pavlovian setup, the dog appeared to be reflexively responding to the tone. However, by employing some ingenious experimental designs, it was evident that the dog was maintaining in memory a history of prior events and using the history to make inferences about what to do next. Subtle changes in the history could cause the dog to show no sign of salivation to the tone at certain times. These and comparable developments led to what has been called the *cognitive revolution in psychology*. The ideas offered here accord with that perspective.

A few remarks about memory are required before discussing how MT generates theory to guide behavior. It is useful to make a distinction between short-term memory (STM) and long-term memory (LTM). The STM is of limited capacity and its contents are continually changing; nothing exists there on a permanent basis. In contrast, the LTM is virtually limitless. It provides a repository of information about events occurring over a lifetime, knowledge needed to understand and speak a language, and all other information available to us from our memory.

It is convenient to think of a memory trace as an array of features with each feature having multiple values (e.g., color as a feature and blue as a possible value). When an event occurs, a trace is activated in STM; the trace serves as a probe to retrieve traces from LTM that are similar to the present event. The MT then uses this information to make an appropriate response. Once the individual responds and an outcome (e.g., a positive or negative reward) has occurred, that information is added to the memory trace. The contents of STM is then stored in LTM. The first occurrence of an event produces a weak memory trace. A repetition of the event produces a new trace in STM; if the new trace (serving as a probe) retrieves the earlier trace, then the two traces can be linked together. Thus, repetitions of the same event can build a rich memory trace that is readily retrieved with an appropriate probe.

Some examples may be helpful. The arrival of a close friend will immediately retrieve a strong memory trace with information about her name, an image, family history and other information that has been linked to the primary trace over repeated exposures. In addition, there may be secondary memory traces that have not been linked to the primary trace; only with a more extended search and a better probe, can they be retrieved. Contrast this example with an individual you met once years ago. You may have several memory traces of her that have never been linked together. When meeting her again, you may retrieve a trace that causes you to recognize her and only later retrieve a trace that lets you recall her name and where you met her.

One more example, a conjecture about my own memory traces for the Pythagorean Theorem. There is probably a primary trace that includes the words "right triangle", an equation "$a^2 + b^2 = c^2$", an image of a right triangle, and the phrase "Greek philosopher Pythagoras". Any one of these entries would serve as a probe to retrieve the memory trace. However, there may be other traces such as an image of a square, each side of length a + b, with its four corners folded as a clue to a proof of the Pythagorean Theorem. Tucked elsewhere in memory might be an image of my high school geometry teacher, Pythagoras' dingdong theory of language, and so forth.

When the conscious mind is not being bombarded by external stimuli, or during certain stages of sleep, a process called *trolling* occurs in LTM. The MT trolls through memory searching for traces that contain similar information. When several similar traces are identified, inspection and manipulation of the group may yield information that was not evident when each trace is examined individually. If reification occurs, the traces can be elaborated and possibly linked together. The trolling process can lead to changes in the traces that make them functionally more accurate, but it can also be the source of distortions, or what is known as *false memories*.

Let me give two hypothetical examples of the trolling process, one exemplifies the *tip-of-the-tongue* phenomenon and the other the *Aha* moment. For the first example, suppose you encounter a former classmate that you have not seen for many years. Immediately, you retrieve a wealth of information about him, but simply cannot recall his name. Some hours later, when you are engaged in something entirely different, the name suddenly pops into consciousness without warning. You were not thinking of him at the time, but suddenly his name appeared. For the Aha example, a young student has tried unsuccessfully to provide a proof for the Pythagorean Theorem. After several hours of effort, she gives up and goes to bed. Once asleep, the trolling process begins to review and manipulate various memory traces of squares and right triangles. One manipulation folds a corner of a square to form a right triangle, and the same fold is repeated on the other three corners. What emerges is an image of a smaller square with its sides formed by the hypotenuse of the four right triangles. Voila, the new image suggests a proof for the theorem and our student is suddenly wide awake (Aha!). We have all had experiences of this sort involving the trolling process operating at a subconscious level.

The formation of a memory trace is prone to error, and its transfer to LTM can lead to additional errors. On later retrievals and during the trolling process different memory traces can combine, with the possibility of modifying the original memory. Further, emotional and motivational states can influence the storage and retrieval processes. Although there is no reliable experimental evidence for Freud's concept of *repressed memories*, there may be traumatic events that cannot be retrieved except under very unusual circumstances.

As information accumulates in LTM, some of it is summarized in symbolic form. Symbols are introduced to represent objects, ideas, and thoughts. These symbols, in turn, form the basis for language. For example, the memory trace for the symbol "chair" would include an image of a chair, the printed word for chair, the sound of the word, and other defining

information. As information accumulates in LTM, the trolling process identifies components embedded in a variety of memory traces that reference chair. These components are then assembled to form a new memory trace that is the symbolic representation of chair. The trace becomes stronger over time as it is more precisely defined by additional inputs. The trolling process—the ability to form new memory traces without external sensory stimulation—is key to understanding many psychological phenomena.

Pattern recognition is an aspect of the trolling process. While trolling, the MT may identify a pattern of events and generate a hypothesis about how these events are related. If the hypothesis is verified by subsequent experience, then it is eligible to be integrated into the mind's theory. As hypotheses accumulate, the MT continues to update and modify the theory. These hypotheses are the building blocks of the mind's theory.

LTM usually is described as of virtually unlimited size. I have often thought that an XPRIZE should be established for anyone who can offer a meaningful estimate of LTM, for example, measured in megabytes. To make such an estimate would require a better understanding of the brain than now exists. Our intuitions about LTM size may be misleading. When I see an old movie that I saw some 70 years ago, I have the feeling that I am aware of every detail of the movie (including dialogue) as it was originally shown. Possibly a series of memory traces are stored in LTM that includes such detail. However, it may be that only gists of the movie are in LTM and the recall process fills in the blanks so that I experience total recall.

Earlier, I suggested that the mind's theory is comparable to scientific theories like those of Newton and Einstein. I need to add a caveat at this point. When a theory is advanced in a scientific field, it must satisfy some basic requirements: the assumptions (axioms) of the theory are logically consistent; the theory explains a range of known phenomena, and makes predictions about new phenomena some of which may have already been confirmed by further research. A

scientific theory might be best thought of as a singular accomplishment of a MT (or a group of MTs) that explains a limited set of phenomena, compared to the virtually limitless tasks the MT must deal with to meet the day-to-day needs of an actual person. However, both have the same goal, to formulate a theory that accounts for past observations and predicts outcomes when faced with new situations.

The MT confronts an ever-changing environment. Its database in memory is error-prone and different memory traces can be contradictory. How the MT interprets the database and what information has priority at any moment can lead to a theory that at times generates inconsistent predictions. Ambiguous situations can create similar problems, because whatever MT predicts will tend to be regarded as correct in the absence of contradictory feedback; research has shown that the mind has a bias to perceive what it expects to perceive. Further, there is the possibility of dissociation. Depending on the particular history of events, components of the theory can function independently at times. An extreme example of dissociation is the clinical case of *dissociative identity disorder*, also known as multiple personality disorder. Notwithstanding these limitations, the MT has proved to be remarkably effective. From a Darwinian perspective, the ability of the MT to predict the future and adapt to change has been key to the success of *Homo sapiens*.

It is evident from our discussion that MT is a complex system and many questions remain, but there is relevant on-going research. In the psychological and cognitive sciences, work on pattern recognition, Bayesian hypothesis testing, problem-solving, and higher-order thinking has identified attributes of a model for MT. In the field of computer science, developments in artificial intelligence and machine learning are relevant. Of special interest, is the work on *deep learning* dealing with computational algorithms and heuristics for speech recognition, computer vision and natural language processing. Building on current research, the

formulation of a general model of the MT soon may be within reach. The model should offer insights into how the MT develops over a lifetime and has changed through the course of evolution.

ENDNOTES

1. To my knowledge, the phrase "The Mind's Theorist" has not been used in the psychological literature. However, the general concept, in various forms, is not new and can be traced back to Plato in the first book of the *Republic*. A particularly elegant realization is due to the British psychologist Kenneth Craik. I quote from his 1943 book *The Nature of Explanation*:

> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it. Most of the greatest advances of modern technology have been instruments which extended the scope of our sense-organs, our brains or our limbs. Such are telescopes and microscopes, wireless, calculating machines, typewriters, motor cars, ships and aeroplanes. Is it not possible, therefore, that our brains themselves utilise comparable mechanisms to achieve the same ends and that these mechanisms can parallel phenomena in the external world as a calculating machine can parallel the development of strains in a bridge?

Craik played an important role in laying the foundation for what I referred to in this essay as the cognitive revolution in psychology. He died in a bicycle accident in 1945 at the young age of 31. A tragic loss for science.

2. For background on the type of memory system described in this essay, see the following references.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence and J. T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 2*, 89-195. \New York: Academic Press.

Atkinson, R. C. and Shiffrin, R.M. (2016). Human Memory: A Proposed System and Its Control Processes. In R.J. Sternberg, S.T. Fiske and D.J. Foss (Eds.), *Scientists Making a Difference*, 115-118. New York: Cambridge University Press.

Nelson, A. B. and Shiffrin, R.M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120(2); 356-394*.

Raaijmakers, J. G. W. and Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In Bower, G. H. (Ed.), *The Psychology of Learning and Motivation, Vol. 14*, 207-262. New York: Academic Press.