A Turing Test for Computational and Associative Theories of Learning

Russell M. Church[1]

Department of Psychology, Brown University, Providence, Rhode Island

Abstract

A Turing test is proposed to evaluate current computational and associative models of learning, and to guide theoretical developments. This test requires a specification of the procedures to which the model applies, a sampling of procedures and response measures, and an objective way to determine the difficulty of discriminating the responses of the model from the responses of the animal. Scalar timing theory is used as an example of a well-developed computational theory of timing that involves addition, multiplication, division, and sampling. The behavioral theory of timing is used as an example of a well-developed associative theory of timing that involves state transitions and strengthening of connections. A Turing test provides a way to evaluate such theories.

Keywords

Turing test; scalar timing theory; behavioral theory of timing

In his *Principles of Psychology*, James (1890) introduced generations of readers to the concepts of both associative mechanisms and mental faculties. Based on his chapters on thinking, association, and habit formation, one could think of individuals as deterministic robots; but based on his chapters on attention, memory, the perception of space, and the sense of time, one could think of individuals as active agents.

Standard textbooks of the 1950s continued to express these two views of the nature of individuals, even within the same chapter. A chapter on animal learning could begin with a description of animals as deterministic machines, giving examples of reflex substitution based on Pavlov's (1927) observations. The same chapter could end with a description of animals as intelligent beings, giving examples of the insight of Koehler's (1925) chimpanzees. Current introductory textbooks do the same.

Quantitative models of associative mechanisms and computational models of animals as information processing agents are now widely available. There has been considerable development of associative models of learning in terms of neural networks (Haykin, 1999; Sutton & Barto, 1998). These are usually implemented as computer programs that contain many simple elements, with rules for the influence of the environment on the elements, for strengthening or weakening connections between the elements, and for the effect of the elements on responses. The essential ideas of associative learning are contained in the neural network models: The main one is that animals learn associations between stimuli, responses, reinforcers, or some combination of these elements. There has also been development of computational models of learning (Gallistel, 1990). The current computational models of learning contain some of the alternative ideas about the learning process: The main one is that animals learn values on various dimensions, such as space, time, number, and rate (e.g., they learn the time from stimulus onset to reinforcement).

At the present time it is not possible to determine which of these approaches will be more productive. Worse still, there is no standard method for evaluating alternative models. In this article, I describe how a Turing test can be applied to this problem.

Some theories of learning are clearly based on associative mechanisms, and others are based on computational mechanisms (Church & Kirkpatrick, 2001). (I use the terms "theory" and "model" interchangeably in this article.) To illustrate the use of a Turing test, I outline two particular theories as they would be applied to a fixed-interval procedure. In this procedure, food is delivered to a rat the first time the rat pushes a lever 1 min or more after the onset of an auditory stimulus. Although the two theories I describe are different in many ways, they have the same input (stimuli and reinforcements) and the same output (responses); and this is the same as the input to the animal (stimuli and reinforcements) and the output from the animal (responses). Thus, it is possible to compare simulated data from alternative models with the observed results of experiments.

SCALAR TIMING THEORY: A COMPUTATIONAL THEORY OF TIMING

Scalar timing theory is an example of a well-developed computational theory of timing that involves addition, multiplication, division, and sampling (Gibbon, Church, & Meck, 1984). According to the theory, timing involves a clock, a memory, and decision processes. The clock consists of a pacemaker, a switch (a logical device), and an accumulator adder. The pacemaker emits pulses according to some mean rate and distribution form; at the onset of the auditory stimulus, the switch between the pacemaker and accumulator closes, and pulses from the pacemaker flow into the accumulator; the sum of pulses in the accumulator is the representation of duration. At the occurrence of a reinforcer, the switch opens, and the flow of pulses into the accumulator stops.

The memory assumed by scalar timing theory is a set of specific examples of the number of pulses at the time of reinforcement. When food is delivered in the fixed-interval procedure, the number of pulses in the accumulator is transferred to memory, multiplied by a memory storage constant. Multiplication by a value other than 1.0 leads to memories of time intervals that are different from those that were perceived. This new information does not affect any of the old, previously stored, information. At the onset of the auditory stimulus, a single random example from memory is sampled; in other words, retrieval from memory requires random sampling.

A decision whether or not to respond is based on a comparison between the number in the accumulator and the number sampled from memory. The process requires the absolute value of the difference between the number in the accumulator and the number in memory, and the ratio of this value to the number in memory. Clearly, scalar timing theory is a computational theory because it involves a logical device, an adder, multiplication, division, determination of an absolute value, and random sampling. It does not use an associative memory structure, but instead relies on a list of examples.

THE BEHAVIORAL THEORY OF TIMING: AN ASSOCIATIVE THEORY

In contrast, the behavioral theory of timing is an example of a well-developed associative theory of timing that involves transitions between states and the strengthening of connections (Killeen & Fetterman, 1988; Machado, 1997). There are three parts of the theory: behavioral states, learned associations, and a response rule.

At the onset of the auditory stimulus in a fixed-interval procedure, a sequence of behavioral states occurs; these states serve as the perceptual representation of time since stimulus onset. Thus, time is not represented by an amount (as in scalar timing theory), but as a specific state. The strength of a state increases when food is delivered, and decreases at all other times; thus, memory consists of strong or weak connections between states and a particular time since stimulus onset. The decision to respond at any given time is based on a response rule that compares the strength of the associative connection with a criterion value. If the strength is greater than the criterion, a response occurs; otherwise, it does not. The behavioral theory of timing is an associative theory because it involves behavioral states that are strengthened by reinforcement and weakened by nonreinforcement.

THE COMPETITIVE APPROACH TO THE EVALUATION OF THEORIES

The computational model (scalar timing theory) and the associative model (behavioral theory of timing) can be used to predict results from the same experimental procedures, such as the fixed-interval procedure.

When theories are compared in terms of how well they fit empirical data, it is typically the case that the supporter for a particular theory determines the procedures and response measures to be used. There are several problems with this competitive approach. First, some theories may be more successful with some procedures and response measures than others. Second, all current theories are known to be flawed. Thus, paradoxically, this method can be used to demonstrate that each theory is better than each other. Finally, each of the models has many assumptions (in the current example, assumptions about perception, memory, and decision processes). If a model successfully accounts for the data of an experiment, it is difficult to know which of its assumptions were necessary, and if a model fails to account for the data of an experiment, it is difficult to know which of its assumptions were responsible.

A TURING TEST FOR THE EVALUATION OF THEORIES

Psychologists have generally accepted the methods of statistical analysis for evaluating factual statements about experimental results, but they have not adopted objective standards for evaluating explanatory statements. A Turing test may provide a way to evaluate current quantitative theories of conditioning and timing, and to guide their development.

Alan Turing introduced what has come to be called the Turing test in an article titled “Computing Machinery and Intelligence,” in the Journal *Mind* in 1950. His problem was to answer the question, Can machines think? He began by considering the normal approach to the question, which is to begin by defining “machine” and “think” as the words are typically used, but he rejected this as the basis of a solution.

He then adopted a behavioristic approach to the problem of whether or not machines think. For this he adapted what he called “the imitation game.” In his version of that game, a programmed computer and a person are in a room separate from an interrogator who attempts to determine which is which. The interrogator asks questions, the computer or the person provides answers, and the interrogator attempts to distinguish between them. Communication, of course, is done in writing.

The Turing test may serve an important function in the evaluation of theories of conditioning and timing. Instead of asking questions, the interrogator chooses some procedure to administer. This procedure consists of stimuli and reinforcements delivered according to some schedule. The same procedure is administered to the animal and to the model (see Fig. 1). For example, the procedure might be the fixed-interval schedule described previously. The output from the model, as well as the output from the animal, is the time of occurrence of each response. Now the problem for the analyst is to classify whether a given time series is one that was produced by a rat or a model. For a theory to pass

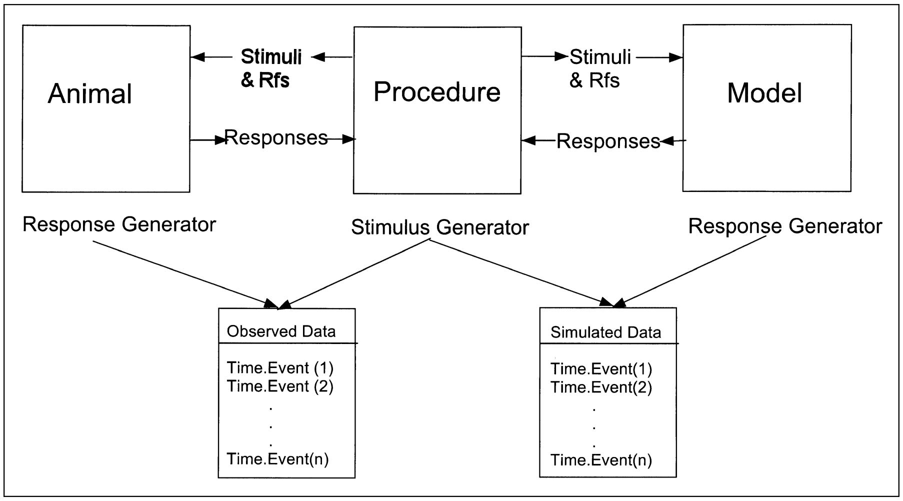

Fig. 1. Diagram of a Turing test for the evaluation of a theoretical model of learning. The procedure administers stimuli and reinforcements (Rfs) to the animal and the model; the animal and model deliver responses to the procedure. The interaction of the procedure and animal generates the observed data; the interaction of the procedure and the model generates the simulated data. Data consist of times of occurrence of stimuli and reinforcements (generated by the procedure) and times of occurrence of responses (generated by the animal or model). The task of the analyst is to determine whether a given set of data was produced by the animal or the model. (From “Quantitative Models of Animal Learning and Cognition,” by R.M. Church, 1997, *Journal of Experimental Psychology: Animal Behavior Processes*, *23*, p. 387. Copyright 1997 by the American Psychological Association. Adapted with permission.)

a Turing test, an informed analyst should not be able to distinguish the output produced by the theory from the output produced by an animal, for any procedure in the domain of the theory.

This is not yet an objective method for testing a model because the conclusions depend on the expertise of the interrogator, the amount of information provided, and a criterion for success. Turing (1950) recognized these problems. He wrote:

> I believe that in about 50 years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an *average* interrogator will not have more than *70 per cent chance* of making the right identification after *five minutes* of questioning. (p. 442, italics added)

Selection of Procedures

Just as a computer program can more easily pass a Turing test of intelligence if the interrogator does not ask challenging questions, a model will be more easily confused with an animal if the procedure is not challenging. The theory may produce output that is indistinguishable from an animal's output in some conditioning and timing procedures, but not in all of them. An objective way to select procedures for testing is needed. If the domain of all possible procedures can be specified, then sampling of the procedures is possible. For example, the domain may be procedures that use rats in a box with stimuli, response alternatives, reinforcements, and contingency specifications. Procedures can be selected from such well-defined sets of procedures on the basis of citation frequency, random sampling of published experiments, or complete census; or a random sample of procedures can be obtained from an organized list of procedures that is prepared.

Selection of Response Measures

For the Turing test of quantitative theories of conditioning and timing, the form of the output of the model is the same as the form of the output from the animal. In both cases, it consists of the time of occurrence of each response. But analysis is typically conducted on various summary statistics. The model may produce output that is indistinguishable from the output of a rat in the case of some summary statistics (such as the mean response frequency), but not other summary statistics (such as the distribution of interresponse time). Some sampling of response measures can be done. Response measures can be selected from well-defined sets of measures on the basis of citation frequency, random sampling, or complete census.

Selection of a Decision Criterion

Finally, it is essential to develop an objective criterion for deciding whether the data came from an animal or from the model. One way to provide an objective interrogation is to use a neural network (Haykin, 1999) rather than a person to make this judgment. An approach is to train the neural network on examples labeled as coming from an animal, and other examples labeled as coming from a model. Then, unlabeled examples may be given, and the neural network attempts to identify the source (animal or model). Such classifiers can be made with limited or considerable power to discriminate between time series generated by different sources. A simple neural network classifier may confuse the output of a model with the output from an animal, but a more complex neural network may distinguish between them. Thus, the complexity of the neural network required to distinguish between the output of the model and the output of the animal may serve as a measure of the goodness of the model.

This Turing test should be done with many different procedures, and with many different descriptive measures. This is a high standard that discourages premature enthusiasm for a flawed theory. Before a serious attempt is made to meet this goal, one cannot know whether or not it is possible.

In the case of the specific Turing test I am suggesting here, it is of course not possible at this time to know if the model that successfully passes the test will be a computational or associative theory of learning. There is the possibility that no psychological model of learning will be able to pass a Turing test because behavior is too complex. The behavior of a rat is determined by billions of synapses, unique experiences, and, some people believe, a little free will. But we have extensive knowledge of the determinants of behavior, both theories and data, and excellent facilities, including behavioral laboratories, computers, and software. Some people believe that a Turing test for the psychology of learning is too high a standard because psychology is a young science, but we have been using that excuse for too long. Research on animal learning has been an active part of the field for more than 100 years (Thorndike, 1898).

CONCLUSIONS

In the past, psychologists have operated without agreed-upon standards for the evaluation of theories. Different theories have been tested using different procedures, different response indices, and different standards of evaluation. A Turing test provides a common ground to evaluate alternative theories, and encourage the development of better ones.

Recommended Reading

Church, R.M. (1997). (See References)
Gibbon, J., Church, R.M., & Meck, W.H. (1984). (See References)
Machado, A. (1997). (See References)
Turing, A.M. (1950). (See References)

Acknowledgments—This article is based on a talk at a symposium on Learning: Association or Computation sponsored by the Department of Psychology, Center for Cognitive Science, Cognitive Development Laboratory, and Laboratory for Language and Cognition at Rutgers University, November 4, 1998. Parts of the article appeared in Brown University's *Faculty Bulletin*, Vol. 11 (November 1998), pp. 33-35. The research used for the development of the ideas expressed in this article was supported by National Institute of Mental Health Grant MH44234 to Brown University.

Note

[1] Address correspondence to Russell M. Church, Department of Psychology, Box 1853, Brown University, Providence, RI 02912; e-mail: russell_church@brown.edu.

References

Church, R.M. (1997). Quantitative models of animal learning and cognition. *Journal of Experimental Psychology: Animal Behavior Processes, 23*, 379–389.

Church, R.M., & Kirkpatrick, K. (2001). Theories of conditioning and timing. In R.R. Mowrer & S.B. Klein (Eds.), *Contemporary learning: Theory and application* (pp. 211–253). Hillsdale, NJ: Erlbaum.

Gallistel, C.R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/MIT Press.

Gibbon, J., Church, R.M., & Meck, W.H. (1984). Scalar timing in memory. In J. Gibbon & L. Allan (Eds.), *Timing and time perception. Annals of the New York Academy of Science*, *423*, 52–77.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

James, W. (1890). *The principles of psychology*. London: Macmillan.

Killeen, P.R., & Fetterman, J.G. (1988). A behavioral theory of timing. *Psychological Review, 95*, 274–295.

Koehler, W. (1925). *The mentality of apes*. New York: Harcourt, Brace.

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review, 104*, 241–265.

Pavlov, I.P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (B.V. Anrep, Ed. & Trans.). London: Oxford University Press.

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

Thorndike, E.L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplement, 2*(4, Whole No. 8), 1–109.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.