

ories, 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.

Special Section

The Four Causes of Behavior

Peter R. Killeen¹

Department of Psychology, Arizona State University, Tempe, Arizona

Abstract

Comprehension of a phenomenon involves identifying its origin, structure, substrate, and function, and representing these factors in some formal system. Aristotle provided a clear specification of these kinds of explanation, which he called efficient causes (triggers), formal causes (models), material causes (substrates or mechanisms), and final causes (functions). In this article, Aristotle's framework is applied to conditioning and the computation-versus-association debate. The critical empirical issue is early versus late

reduction of information to disposition. Automata theory provides a grammar for models of conditioning and information processing in which that constraint can be represented.

Keywords

associations; automata; causality; explanation; models

Judging whether learning is better explained as an associative or computational process requires that we clarify the key terms. This essay provides a framework for discussing *explanation*, *association*,

and *computation*; it leaves *learning* as an unexamined primitive.

ARISTOTLE'S FOUR CAUSES

Aristotle (trans. 1929) described four kinds of explanation. Because of mistranslation and misinterpretation by "learned babblers" (Santayana, 1957, p. 238), his four "because [*aitia*]" were derogated as an incoherent treatment of causality (Hocutt, 1974). Although ancient, Aristotle's four (be)causes provide an invaluable framework for modern scientific explanation, and in particular for resolution of the current debate about learning.

In Aristotle's framework, *efficient causes* are triggers, events that bring about an "effect." This is the contemporary meaning of *cause*. Philosophers such as Hume, Mill, and Mackie have clarified the criteria for identifying various efficient causal relations (e.g., necessity, sufficiency, insufficient but necessary

events in the context of otherwise sufficient events). Efficient causes identify the early parts of a sequence that are essential for the later parts; they tell us what initiates a change of state. Jachmann and van den Assem's (1996) "causal ethological analysis" of the courtship behavior of a wasp exemplifies this meaning of *cause*.

Material causes are substrates. These are the most common kinds of causal explanation in use today, exemplified by most of neuroscience and brain-imaging research. Once the machinery has been identified, many people consider the phenomenon explained. Exclusive focus on machinery is known as *reductionism*.

Formal causes are models. Newton's great achievement was to give credibility to such models absent material causes: For him, there were no "hooks and eyes" to gravity—"Hypothesis [concerning underlying mechanism] is no part of my design"—just naked math. This was a difficult position for Newton to adopt, for as a mechanical philosopher he abhorred occult (and thus ad hoc) accounts. Newton would gladly have equipped his theory with hooks and eyes—material causes—but could devise none sufficient to hold the planets in their orbits.

Formal causes are logical maps. Aristotle's favorite form was the syllogism, just as the modern physicist's favorites are differential equations. Such equations describe the course of change from one state to another; in concert with initial conditions (efficient causes), they describe the complete trajectory of change.

No matter how successful formal models are, they are not machines: Mathematical equations describe the trajectories of baseballs and planets, but those bodies do not solve equations to project their moves. The formal models of the contributors to this Special Section are mute concerning efficient cause, substrate, and function. It is

possible to speculate about underlying mechanisms, and to generate formal models of them; but without direct data on those mechanisms, the models are unverifiable conjectures and typically subject to change as fads come and go—they are occult.

Final causes are functional explanations. "To recognize an actual machine, we have to have some idea of what it is supposed to do" (Minsky, 1967, p. 4). Questions like "What is it for?" and "Why does it do that?" call for functional (final) causes; survival of the fittest, optimal foraging theory, and purposive explanations in general provide relevant answers. Most of modern physics can be written in terms of functions that optimize certain variables, such as energy. All laws stated in terms of such optima concern final causes. Common examples are light rays following paths that minimize transit times, animals behaving in ways that maximize genetic representation in succeeding generations, and humans behaving in ways that maximize the benefits for a population. Final causes were given a bad name (*teleology*) because they were treated as errant formal, material, or efficient causes. A reason giraffes have long necks is to let them browse high foliage; this final cause does not displace formal (variation and natural selection) and material (genetic) explanations; nor is it an efficient cause (Lamarckianism). But none of those other causal explanations make sense without specification of the final cause. Biologists reintroduced final causes under the euphemism "ultimate mechanisms," referring to the efficient and material causes of a behavior as "proximate mechanisms."

Two systems that share similar final causes may have quite dissimilar substrates. Analyses of evolutionary analogues—such as wings in insects, birds, and bats—provide useful functional information (concerning, e.g., convergent evolution-

ary pressures and varieties of strategies adequate for that function), even though the wings are not homologues (i.e., are not evolved from the same organ in an ancient forbear). Analogical-functional analyses fall victim to "the analogical fallacy" only when it is assumed that similarity of function entails similarity of efficient (evolutionary history) or material (physiological) causes. Such confounds can be prevented by accounting for each type of cause separately.

Efficient causes, then, are the initial conditions for a change of state; final causes are the terminal conditions; formal causes are models of transition between the initial and terminal conditions; material causes are the substrate on which these other causes act.

EXPLAINING CONDITIONING

Skinner (1950) railed against formal ("theorizing"), material ("neuro-reductive"), and final ("purposive") causes, and scientized efficient causes as "the variables of which behavior is a function." He was concerned that complementary causes would be used in lieu of, rather than along with, his functional analysis. But of all behavioral phenomena, conditioning is the one least able to be comprehended without reference to all four causes: The ability to be conditioned has evolved because of the advantage it confers in exploiting efficient causal relations.

Final Causes

Conditioning shapes behavioral trajectories into shortest paths to reinforcement (Killeen, 1989). When a stimulus predicts a biologically significant event (an unconditioned stimulus, US), animals improve their fitness by "learning associations" among external

events, and between those events and appropriate actions. Stable niches—those inhabited by most plants, animals, and fungi—neither require nor support learning: Tropisms, taxes, and simple reflexes adequately match the quotidian regularities of light, tide, and season. However, when the environment changes, it is the role of learning to rewire the machinery to exploit the new contingencies. Better exploiters are better represented in the next generation. This is the final—ultimate, in the biologists' terms—cause of conditioning. Understanding learning requires knowing what the learned responses may have accomplished in the environments that selected for them.

Efficient Causes

These are the prototypical kinds of causes, important enough for survival that many animals have evolved sensitivity to them. Parameters that are indicators of efficient causes—contiguity in space and time, temporal priority, regularity of association, and similarity—affect both judgments of causality by humans (Allan, 1993) and speed of conditioning (Miller & Matute, 1996).

Material Causes

The substrate of learning is the nervous system, which provides an embarrassment of riches in mechanisms. Development of formal and efficient explanations of conditioning can guide the search for operative neural mechanisms. In turn, elucidation of that neural architecture can guide formal modeling, such as parallel connectionist models—neural nets—that emulate various brain functions. Each of the four causes is a resource for understanding the others.

Formal Causes

Models are proper subsets of all that can be said in a modeling language. Associationist and computational models of learning are formulated in the languages of probability and automata, respectively. Their structures are sketched next.

Associative Models

Material implication, the *sufficient* relation (if C , then E ; symbolized as $C \rightarrow E$), provides a simplistic model of both efficient causality and conditioning. It holds that whenever C , then also E ; it fails whenever C and $\sim E$. When the presence of a cue (C , the conditioned stimulus, or CS) accurately predicts a reinforcer (E , the US), the strength of the relation $C \rightarrow E$ increases. The conditional probability of the US given the CS — $p(E|C)$ —generalizes this all-or-none relation to a probability. Animals are also sensitive to the presence of the US in the absence of the CS , $p(E|\sim C)$; only if this probability is zero is a cause said to be *necessary* for the effect. Unnecessary effects degrade conditioning, just as unexpected events make an observer question his grasp of a situation.

Good predictors of the strength of learning are (a) the difference between these two conditional probabilities and (b) the *diagnosticity* of the CS , $p(E|C)/p(E)$, which is the degree to which the cause (CS) reduces uncertainty concerning the occurrence of the effect (US). As is the case for all probabilities, measurement of these conditionals requires a defining context. This may comprise combinations of cues, physical surroundings, and history of reinforcement. Reinforcement engenders an updating of the conditionals; speed of conditioning depends on the implicit weight of evidence vested in the prior conditionals. The databases for some con-

ditionals—such as the probability of becoming ill after experiencing a particular taste—often start small, so that one or two pairings greatly increase the conditional probability and generate taste aversions. Earlier pairings of the taste and health, however, will give the prior conditionals more inertia, causing the conditional probability to increase more slowly, and possibly protecting the individual from a taste aversion caused by subsequent association of the taste with illness. More common stimuli, such as shapes, may be slow to condition because of a history of exposure that is not associated with illness. Bayes's theorem provides a formal model for this process of updating conditional probabilities. This exemplifies how subsets of probability theory can serve as a formal model for association theory. Associative theories continue to evolve in light of experiments manipulating contextual variables; Hall (1991) provided an excellent history of the progressive constraint of associative models by data.

Computational Models

Computers are machines that associate addresses with contents (i.e., they go to a file specified by an address and retrieve either a datum or an instruction). Not only do computers associate, but associations compute: "Every finite-state machine is equivalent to, and can be 'simulated' by, some neural net" (Minsky, 1967, p. 55). Computers can instantiate all of the associative models of conditioning, and their inverses. For the computational metaphor to become a model, it must be restricted to a proper subset of what computers can do; one way to accomplish this is via the theory of *automata* (Hopkins & Moss, 1976). Automata theory is a formal characterization of computational architectures. A critical distinction among automata is

memory: Finite automata can distinguish only those inputs (histories of conditioning) that can be represented in their finite internal memory. Representation may be incrementally extended with external memory in the form of push-down stores, finite rewritable disks, or infinite tapes. These amplified architectures correspond to Chomsky's (1959/1963) context-free grammars, context-sensitive grammars, and universal Turing machines, respectively. Turing machines are models of the architecture of a general-purpose computer that can compute all expressions that are computable by any machine. The architecture of a Turing machine is deceptively simple, given its universal power; it is access to a potentially infinite memory "tape" that gives it this power. Personal computers are in principle Turing machines, silicon instruments whose universality has displaced most of the brass instruments of an earlier psychology.

The Crucial Distinction

Memory is also what divides the associative from the computational approaches. Early reduction of memory to disposition requires fewer memory states than late reduction and permits faster—reflexive—responses; late reduction is more flexible and "intelligent." Animals' behavior may reflect computation at any level up to, but not exceeding, their memory capacity. Most human behaviors are simple reflexes corresponding to finite automata. Even the most complicated repertoires can become "automatized" by practice, reducing an originally computation-intense response—a child's attempts to tie a shoe—to a mindless habit. The adaptation permitted by learning would come at too great a price if it did not eventually lead to automatic and thus fast responsivity.

Consciousness of action permits adaptation, unconsciousness permits speed.

In traditional associative theory, information is reduced to a potential for action ("strength" of association between the CS and US) and stored on a real-time basis. Such finite automata with limited memories are inadequate as models of conditioning because "the nature of the representation can change—the sort of information it holds can be influenced by [various post hoc operations]" (Hall, 1991, p. 67). Rats have memorial access to more of the history of the environment and consequences than captured by simple Bayesian updating of dispositions. Miller (e.g., Blaisdell, Bristol, Gunther, & Miller, 1998; see also this issue) provided one computational model that exemplified such late reduction.

If traditional associators are too simple to be a viable model of conditioning, unrestricted computers (universal Turing machines) are too smart. Our finite memory stores fall somewhere in between. Automata theory provides a grammar for models that range from simple switches and reflexes, through complex conditional associations, to adaptive systems that modify their software as they learn. The increased memory this requires is sometimes internal, and sometimes external—found in marks, memoranda, and behavior ("gesturing facilitates the production of fluent speech by affecting the ease or difficulty of retrieving words from lexical memory," Krauss, 1998, p. 58). Context is often more than a cue for memory—it constitutes a detailed, content-addressable form of storage located where it is most likely to be needed. Perhaps more often than we realize, the medium *is* memory.

The difference between associativistic and computational models reduces to which automata they are isomorphic with; and this is

correlated with early versus late reduction of information to action. The challenge now is to identify the class and capacity of automata that are necessary to describe the capacities of a species, and the architecture of associations within such automata that suffice to describe the behavior of individuals as they progress through conditioning.

Comprehending Explanation

Many scientific controversies stem not so much from differences in understanding a phenomenon as from differences in understanding explanation: expecting one type of explanation to do the work of other types, and objecting when other scientists do the same. Exclusive focus on final causes is derided as teleological, on material causes as reductionistic, on efficient causes as mechanistic, and on formal causes as "theorizing." But respect for the importance of each type of explanation, and the correct positioning of constructs within appropriate empirical domains, resolves many controversies. For example, associations are formal constructs; they are not located in the organism, but in our probability tables or computers, and only emulate connections formed in the brain, and contingencies found in the interface of behavior and environment. Final causes are not time-reversed efficient causes. Only one type of explanation is advanced when we determine the parts of the brain that are active during conditioning. Provision of one explanation does not reduce the need for the other types. Functional causes are not alternatives to efficient causes, but completions of them.

Formal analysis requires a language, and models must be a proper subset of that language. The signal issue in the formal analysis of conditioning is not association

versus computation, but rather the circumstances of early versus late information reduction, and the role of context—both as a retrieval cue and as memory itself. Automata theory provides a language that can support appropriate subsets of machines to model these processes, from simple association up to the most complex human repertoires.

Comprehension is a four-footed beast; it advances only with the progress of each type of explanation, and moves most gracefully when those explanations are coordinated. It is a human activity, and is itself susceptible to Aristotle's quadripartite analyses. In this article, I have focused on the formal analysis of explanation, and formal explanations of conditioning. Comprehension will be achieved as such formal causes become coordinated with material (brain states), efficient (effective contexts), and final (evolutionary) explanations of behavior.

Recommended Reading

Miller, R.R., Barnet, R.C., & Grahame, N.J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin*, 117, 363–386.

Uttal, W. (1998). *Toward a new behaviorism: The case against perceptual reductionism*. Mahwah, NJ: Erlbaum.

Wasserman, E.A. (1993). Comparative cognition: Toward a general understanding of cognition in behavior. *Psychological Science*, 4, 156–161.

Wasserman, E.A., & Miller, R.R. (1997). What's elementary about associative learning? *Annual Review of Psychology*, 48, 573–607.

Young, M.E. (1995). On the origin of personal causal theories. *Psychonomic Bulletin & Review*, 2, 83–104.

Acknowledgments—This article was written with the support of National Science Foundation Grant IBN 9408022 and National Institute of Mental Health Grant K05 MH01293.

Note

1. Address correspondence to Peter Killeen, Department of Psychology, Arizona State University, Tempe, AZ 85287-1104; e-mail: killeen@asu.edu.

References

Allan, L.G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, 114, 435–448.

Aristotle. (1929). *The physics* (Vol. 1; P.H. Wicksteed & F.M. Cornford, Trans.). London: Heinemann.

Blaisdell, A., Bristol, A., Gunther, L., & Miller, R. (1998). Overshadowing and latent inhibition counteract each other: Support for the comparator hypothesis. *Journal of Experimental Psychology: Animal Behavior Processes*, 24, 335–351.

Chomsky, N. (1963). On certain formal properties of grammars. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Readings in mathematical psychology* (Vol. 2, pp. 125–155). New York: Wiley. (Original work published 1959)

Hall, G. (1991). *Perceptual and associative learning*. Oxford, England: Clarendon Press.

Hocutt, M. (1974). Aristotle's four because. *Philosophy*, 49, 385–399.

Hopkins, D., & Moss, B. (1976). *Automata*. New York: North Holland.

Jachmann, F., & van den Assem, J. (1996). A causal ethological analysis of the courtship behavior of an insect (the parasitic wasp *Nasonia vitripennis*, hym., pteromalidae). *Behaviour*, 133, 1051–1075.

Killeen, P.R. (1989). Behavior as a trajectory through a field of attractors. In J.R. Brink & C.R. Haden (Eds.), *The computer and the brain: Perspectives on human and artificial intelligence* (pp. 53–82). Amsterdam: Elsevier.

Krauss, R. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54–60.

Miller, R.R., & Matute, H. (1996). Animal analogues of causal judgment. In D.R. Shanks, D.L. Medin, & K.J. Holyoak (Eds.), *Causal learning* (pp. 133–166). San Diego: Academic Press.

Minsky, M. (1967). *Computation: Finite and infinite machines*. Englewood Cliffs, NJ: Prentice-Hall.

Santayana, G. (1957). *Dialogues in limbo*. Ann Arbor: University of Michigan Press.

Skinner, B.F. (1950). Are theories of learning necessary? *Psychological Review*, 57, 193–216.