J. EXPT. THEOR. ARTIF. INTELL. 8(1996)399–417

Developmental aspects of expertise: Rationality and generalization

MARK H. BICKHARD

Department of Psychology, Chandler-Ullmann Hall, 17 Lehigh University, Bethlehem, PA 18015, USA
mhb0@lehigh.edu

ROBERT L. CAMPBELL

Department of Psychology, Brackett Hall 410A, Clemson University, Clemson, SC 29634–1511, USA
campber@clemson.edu

Abstract. Successful attempts to explain expertise in human beings, or to capture its properties in expert systems, will have to contend with issues of rationality and generalization. Rationality and generalization pose enough difficulties on a purely synchronic basis. But an account of expertise must be diachronic—it must account for the *development* of rationality and generalization, even in those who are already experts. We describe the obstacles in the path of standard approaches to rationality and generalization, and present an alternative, interactivist treatment of rationality and its development (space forbids us to do likewise for generalization). In the interactivist account, rationality cannot be defined in general as adherence to the rules of a system of formal logic; we propose instead that rationality be understood in terms of the development of negative knowledge—knowing what kinds of errors to avoid. We examine the development of negative knowledge using examples from the history of science, and consider the consequences of an orientation towards negative knowledge for classroom instruction as well as the development of expert systems.

1. The development of expertise

Psychologists and computer scientists pursue different and sometimes conflicting agendas when it comes to expertise. But there are a couple of basic issues on which consensus shouldn't be too hard to find, whatever else we may think is necessary for expertise. First, every human expert (and, consequently, every expert system) must engage in rational problem solving. Second, every human expert (and expert system) must be able to generalize from past problems and their old solutions to future problems and their new solutions.

There are powerful tools now to aid us in modelling capabilities for rationality and generalization, and in designing systems that manifest such capabilities. Some of these tools are described at length elsewhere in this special issue, notably in the articles by Hewett (1996), Leake (1996), and Prietula *et al.* (1996). While such techniques as case-based reasoning are already being put to use in impressive ways, their ultimate adequacy remains a subject for debate (Bickhard and Campbell in press; Bringsjord and Bringsjord 1996; Whitley 1996).

0952–813X/96 $12·00 © 1996 Taylor & Francis Ltd

Our focus here is not on the rational problem solving that experts can already do, or the generalizations that they can already make. Rather, our concern is with the way experts get more expert. How do they become *more* rational? How are they able to generalize in *new* ways?

In other words, we are not just asking a cognitive science question about expertise, or an Artificial Intelligence question—we are asking a developmental question. We will argue that cognitive science and AI have not provided us with the tools we need to model the *development* of expertise in human beings, or to design *developmental* capabilities into expert systems. (Empirical methods for studying the development of expertise raise questions of their own, which are the topic of another article—Campbell and Di Bello 1996.) Our claim is that the development of rationality, and the development of generalization, are beyond the scope of current theories in psychology and current approaches to expert system design. Which is not to say that they are out of reach altogether, just that rather different kinds of thinking will be called for. If our analysis is correct, the developmental problems that we discuss will impose constraints even on those cognitive architectures that are intended only to cover the non-developmental aspects of expertise.

1.1. *The development of rationality*

Let's begin with the question of rationality. It has always been attractive to describe rationality in terms of adherence to the rules laid out in some system of formal logic. In fact, a traditional view held that the laws of logic were the 'laws of thought'. Yet seemingly straightforward definitions of rationality in terms of formal logic would require humanly impossible feats; to qualify as rational, you would have to know everything that follows via valid deductive inference from what you know—or from what you are merely supposing (Cherniak 1986).

Definitions of rationality as adherence to rules of logic would at least have to be seriously restrained to be realistic. And a restrained definition, such as the 'minimal rationality' condition proposed by Cherniak (1986), still requires formal logic to be an adequate description of the constraints that are applied in human thinking (albeit one that needs supplementing with accounts of how extensively we can apply those constraints under time pressure and with limited capacities). But systems of formal logic are at best a partial, higher-level description of those constraints (Brown 1988, Campbell and Bickhard 1986, Hooker 1995); no one can claim to have formalized them all. Indeed, to what extent human beings actually use rules of formal logic in the course of reasoning is hotly debated at present; some cognitive psychologists (Braine and Rumain 1983) maintain that 'mental logic' is widespread in real human thought, whereas others doubt its use almost entirely (Johnson-Laird 1983).

We find it plausible that people sometimes do use rules of logical inference in their thinking (see, for instance, Campbell 1991). But whatever the merits of a 'mental logic' account of specific kinds of thinking at specific points in human development, it can't be telling the whole story about human reasoning. When broader historical and developmental trends are taken into account, it becomes clear why rationality cannot be simply assimilated to logicality. A system of formal logic already contains all of its valid theorems. Logical systems don't grow or become more powerful. Logical systems can't construct new logical systems more powerful than themselves. Yet the history of logic shows that human knowledge of logic and logical systems has developed (Bochenski 1970, Kneale and Kneale 1986) and that more and more powerful systems of logic have been discovered. If rationality simply means following the rules contained within a particular logical system, then the history of logic can't be rational!

Similarly, if being rational means being logical, the development of each individual's knowledge of logic can't be a rational process either (Fodor 1972, Campbell and Bickhard 1987, Bickhard 1991b). Obviously, something is deeply wrong with such conclusions. Yet an appropriate alternative conception of rationality doesn't immediately suggest itself.

We will propose our own solution to the conundrum below. Whether our particular conception of rationality is found acceptable or not, the need for such a solution is pressing. Insofar as models of expertise, or expert systems, presume, implicitly or explicitly, that rationality can be reduced to following rules of formal logic, they cannot be adequate. At best, such models will be radically incomplete, and only an alternative model of rationality will enable us to avoid this incompleteness.

2. Generalization

2.1. *Topologies*

Just as developmental and historical arguments caution us against equating rationality with logicality, they provide counterexamples against standard approaches to generalization. To appreciate this point requires a preliminary formal treatment of generalization.

Generalization can't take place without processes that differentiate similarity and difference. Generalizing from one case to another requires some sense that the two cases are relevantly similar. Similarity, in turn, is closeness or distance on some aspect or measure. For instance, two objects might be similar in colour but not size, or in friendliness but not in intelligence.

Formally, issues of nearness or distance are capturable in topological terms. In a topology, nearness amounts to being in the same (open) sets together (Hocking and Young 1961). True, there are mathematical forms suited to more structured relations of similarity—for instance, uniformities (James 1987) and metric spaces (Geroch 1985). But for the points we need to make here, general topologies will do.

Approaches that are widely relied on in cognitive science and expert system design, such as instance or exemplar-based categorization (Medin and Schaffer 1978) and case-based reasoning (Hammond 1989, Kolodner and Simpson 1989, Riesbeck and Schank 1989, Rich and Knight 1991) plainly need topologies. Without a topology, there is no way to generalize to any cases or instances in preference to any others.

Models of similarity judgments, and designs for processes that must deal with similarity, most often use feature spaces (the many examples include the case-based expert system work of Leake 1996, and Hewett 1996). In a feature space, nearness is a function of the overlap (or lack of overlap) between the features of the items being compared. Features are usually discrete, and some sort of distance measure is usually derived from the feature sets, so as a rule feature spaces are discrete metric spaces.

Many challenges can be brought against feature-space models. Can all similarity relationships be captured with discrete models? Do all processes that are sensitive to similarity have the full mathematical structure that metric spaces require? What about evidence that similarity judgments don't always behave as though similarity is a distance measure? Similarity judgments are in many cases directional, not symmetric; moreover, they are often affected by the context in which they are made (Tversky 1977, Shanon 1988, Medin *et al.* 1993). Simply making a similarity judgment can alter the processes for making similarity judgments (Bickhard and Campbell in press, Gentner and Grudin 1985, Gentner and Rattermann 1991, Gentner and Jeziorski 1994, Gentner and Markman 1995, Medin *et al.* 1993).

There are plenty of technical difficulties for standard models of generalization. But those we have mentioned so far are synchronic difficulties. The challenge is how to capture directionality, context-sensitivity, and alteration through use for particular similarity-sensitive processes within given spaces of representations.

We want to emphasize a different challenge—a diachronic one. No matter how successful discrete metric approaches may be at predicting the results of particular experiments in cognitive psychology or providing a basis for usable expert systems, they depend on a representation space and a feature space that have already been designed and encoded into the system that makes the similarity judgments. There is no way to get new representations, or spaces of representations. Of course, there is no way to get *topologies* on new representational spaces; what's worse, there's no way to get new topologies on the old representational spaces either. Feature-based models cannot account for the emergence of new basic features (or of any other grounds for similarity).

Yet we know that such emergence happens. New kinds of representations get constructed. New relevant similarities get learned within representational spaces. There is ample evidence of such changes, historically and within the development of the individual. Becoming skilled in any new problem domain, like high school algebra, means learning new representational spaces and new kinds of similarities within those spaces—that is, new topologies on them. A basic part of becoming expert in such a domain is recognizing that this algebra problem, or calculus problem, or diplomatic problem, is similar to some other more familiar problem. The representations, and the topologies of similarity, that are involved in expertise are learned. They are neither predesigned nor pre-encoded.

2.2. *New representations?*

Standard approaches to AI and cognitive science can't account for the construction of genuinely new representations. Jerry Fodor (1975, 1981) has become notorious for making this very point, although for the most part cognitive scientists still don't take him seriously. Basically, standard approaches can account for new organizations of pre-existing representational elements (features, semantic primitives, and so forth), but they cannot account for the origins of those supposed elements, or for the emergence of representation out of non-representational phenomena (Bickhard 1991b, 1993, Bickhard and Terveen 1995, Campbell and Bickhard 1987). Yet, on a cosmological scale, mental representation came into being at one time or another: there was no representation at the Big Bang, and there is representation now. Most developmental psychologists believe that representation also comes into being during the course of each person's life. No model of cognition is truly adequate unless it can account for representational emergence.

2.3. *New topologies?*

Some in AI and cognitive science are aware that their theories make no provision for emergent representation. Hardly anyone recognizes that such theories cannot account for emergent topologies. Until quite recently, it was thought that encoded features, whatever the difficulties in explaining where *they* came from, would be a powerful enough resource. The problem of new topologies is only now coming to the fore (Medin *et al.* 1993, Bickhard and Campbell in press).

For a standard feature-based model, creating a new topology in a new representational space is partly a matter of creating new representations—to wit, new encoded features. For new topologies to emerge, however, it would also be necessary to appropriately associate the new encoded features with every new representational element in the new space. Learning which features belong with which elements could be very difficult. It might even be computationally intractable. Indeed, to the extent that the new features represent conditions under which changes occur or don't occur, this task of associating features and representational elements becomes a frame problem. We know that frame problems (Pylyshyn 1987, Ford and Hayes 1991, Bickhard and Terveen 1995, Ford and Pylyshyn in press) are computationally intractable.

Even if the features in the new spaces could be associated with the right representational elements, that would still not be enough for similarity-sensitive processes. To be sensitive to similarity, the system would need more than the features and their associations with new representational elements. It would also need programs to calculate the feature-based distances and influence further processing in the system according to those distances. Where would such programs come from?

It gets worse. For the most part, new representational spaces won't get constructed *in toto*, because most of them will be too large, and some of them will be unbounded. How large, for instance, is the space of representations for polynomials in high school algebra? If the entire space had to be constructed before similarity judgments could be made, virtually no one would be able to recognize the degree of similarity between one polynomial and another! (See Bickhard 1993, and Bickhard and Terveen 1995, for further examples of unbounded topologies.) Features cannot be associated with new representational elements via element-by-element constructions, because most new representations will be constructive *potentials* of the system rather than preconstructed elements already stored within the system. The knowledge that is embodied in topologies will be difficult to capture in the explicit representations characteristic of expert systems, because much of it is implicit.

2.4. *Old problems and new solutions?*

Processes that are to be sensitive to new topologies must be sensitive to representations that are already constructed, such as old problems that have already been solved and their solutions, as well as to representations yet to be constructed, like new problems and their possible solutions. Moreover, the old problem representations and solutions must get privileged treatment. Because they are already known to work, they are needed to serve as anchors for similarity comparisons.

The need to accord a privileged status to old, successful constructions seems to face us with a dilemma. What if topological information is somehow constructed when new representations are, and both of the representations whose similarity is to be judged are constructed in the moment? Then there would be no way to distinguish between old, solved problems and new problems, because both kinds of representations would be newly constructed. What if old problems don't have to be newly constructed, because they have been stored in some memory in the past? Then the constructive processes must manifest a topological sensitivity to the old representations in the process of constructing the new representations—in a manner that will distinguish between solved problems and new problems.

Now there are ways of building at least some of the needed characteristics into one's cognitive architecture. If the construction of new representations involved the construction of new feature encodings that were correctly associated with them, *and* the topologically sensitive processes in the system set a bit to keep track of what came from the 'new construction register' and what was retrieved from the 'old problem and solution register', then the topological information, including the proper asymmetry between old and new, would be technically available. But it would be available only because it was designed into the system or built into the model *a priori*. And when we fall back on predesigned architectures for similarity and generalization, we forfeit any opportunity to model the way they are learned, or the way they develop.

Accounting for generalization is a multilayered problem. The interactivist treatment of generalization, similarity, and topologies can't be squeezed into the usual confines of a journal article (for an in-depth presentation, see Bickhard and Campbell, in press). In any case, the details of this treatment are not needed to make our general point. Any attempt to account for the development and learning of new generalization abilities must meet strong new requirements, requirements whose very existence usually goes unacknowledged. And if the tools available to symbol-manipulation approaches, or connectionist approaches, or to any of the prevailing conceptions of encoded mental representation, are inadequate to account for diachronic phenomena, suspicion naturally arises that they aren't fully adequate to account for their chosen range of synchronic phenomena either. If a model M of phenomenon P makes the development of P impossible, then M can't be a correct model of P. After all, P can't exist if it can't come into existence.

The constraints imposed by the need to account for learning and development haven't been widely appreciated (Bickhard 1979, 1991b, Campbell and Bickhard 1987, Terwilliger 1968). Up to now they have intruded very little into discussions of expertise, or of rationality and generalization more broadly.

3. Rationality and critical principles

An expert is someone who knows some of the worst mistakes that can be made in his subject, and how to avoid them. Werner Heisenberg

The development of generalization poses too many complex problems to be tractable within the available space. A conception of rationality that addresses the diachronic issues is a lot easier to sketch. In place of the conventional synchronic assumption, that rationality means logicality, we will need two diachronic ones: (1) knowledge must be constructed; and (2) human beings are capable of epistemic reflection. Let's begin with constructivism.

3.1. *Constructivism*

3.1.1. *The demise of empiricism*. Constructivism is directly opposed to the classic empiricist doctrine that no construction is required because the world impresses itself onto the mind. Aristotle (1941) maintained that the environment impresses various forms onto the mind, in the same fashion that a ring leaves an impression in a blob of sealing wax. Although empiricist ideas have undergone 2000 years of evolution in the interim, contemporary appeals to transduction as a source of sensory knowledge, and to induction as a source of conceptual knowledge or of scientific generalizations, still draw on the same intuition. Impressed or transduced states are in correspondence with whatever did the impressing, or whatever got transduced. And they are thought to be mental representations by virtue of those correspondences. Activity in the retina, for instance, is supposed to be in correspondence with properties of the light; in consequence, the retinal activity is supposed to represent those properties (Fodor and Pylyshyn 1981).

Despite long and ingenious labour, attempts to make good on empiricist accounts of knowledge have never worked (Suppe 1977). Even if empiricism could explain how we know what is actually the case, it can't explain how we know what is possible or what is necessary. There are nine planets in the solar system, but there is no mathematical necessity that there be nine planets. Whereas it is mathematically necessary that three times three equal nine. Tallying pebbles or carrying out some other kind of empirical observation will show that three times three has always equalled nine whenever data were collected—not that three times three must equal nine (Hume 1739/1888, Harré, 1970, Mackie 1985, Moser 1987).

Empiricism can't even account for knowledge of what is actually happening in the external world. The best that empiricist models can do is show how the world has some causal or functional effect on the knowing agent. It has proven impossible to account for any *knowledge* of the world on that basis (Dancy 1985, Fodor 1987, 1990a, 1990b, Loewer and Rey 1991). Clearly there's a good deal more to say about empiricist epistemologies, but we will proceed on the firm assumption that they have failed (Bickhard 1993, Bickhard and Terveen 1995).

3.1.2. *Knowledge in system organization*. What usually prevents psychologists and computer scientists from moving beyond empiricist assumptions is an apparent lack of alternatives. If knowledge is not impressed by the world into the agent, and constituted as correspondences between the impressing environmental conditions and the impressed mental states, what else could it be? Yet nowadays it's generally admitted that skill—knowing how to bring about various results—is a form of knowledge. Of course, accounts of 'procedural knowledge' in cognitive science and artificial intelligence (Anderson, 1983) normally take 'knowing that' as their standard, rather than 'knowing how.' So they end up reducing skill to encoded propositions, or encoded rules, or some other kind of representation by correspondence. The fundamental alternative to empiricism regards skill—our capacity for interacting successfully with the world—as the basic form of knowledge (Bickhard 1980b 1993, Bickhard and Terveen 1995, Clark 1993, Dreyfus 1967, 1982, 1991).

If knowledge is interactive skill, then knowledge is constituted in the overall organization of the system that is interacting with the world. Knowledge does not take the form of encoded data structures that correspond to structures in the environment. Nor is there any temptation to suppose that knowledge as skill might be impressed into the system. System organization is not in correspondence with aspects of the world; it is interactively competent in the world. System organization can't be impressed; it has to be constructed.

3.1.3. *Evolutionary epistemology*. By and large, knowledge construction is heuristically guided, which means it must involve prior knowledge of the heuristics that might be used and of their value. But models of knowledge and its construction can't logically *require* prior heuristic knowledge, because the origins of that heuristic knowledge must also be accounted for. If all knowledge logically required prior knowledge, then knowledge could never have emerged in the first place. At the logical limit, knowledge construction has to be blind; it must be done without any foreknowledge of what is worth constructing. All the system can do is try a possible organization to find out whether it works; those organizations that enable the system to reach its goals are retained, and those that fail to do so are rejected. And if one is eliminated by selection pressures, then another system organization must be constructed and tried out. Thus, if knowledge is interactive system organization, it must be constructed by variation and selection. Interactivism leads straight to evolutionary epistemology (Bartley 1987, D. Campbell 1974, Hahlweg and Hooker 1989, Hooker 1995, Wuketits 1990).

3.1.4. *Recursivity and metarecursivity*. Within a general interactivist framework, knowledge construction can take on further properties that are of developmental significance. First, it can be *recursive*. That means that previous constructions can be used in the service of further constructions. For instance, previously successful constructions could be used as units in new constructive trials. The system doesn't have to reinvent the wheel every time a wheel might be needed.

Second, there can be *meta*recursive construction. Metarecursive construction operates on the construction processes themselves, rather than their products. A system that is capable of metarecursive construction can learn how to learn better. It can get better at its constructive processes.

3.2. *Epistemic reflection*

Besides interactive representation, and a variation and selection process with recursive and metarecursive capabilities, our account of rationality requires epistemic reflection. By epistemic reflection, we mean knowledge and thought *about* knowledge and thought. In other words, reflective consciousness. In the early 20th century, anti-ontological empiricism drove consciousness completely out of the field of psychological inquiry. It is only in the last twenty years that consciousness has begun to regain its status in philosophy and in cognitive science (Baars 1986, Dennett 1991, Mathews *et al.* 1996).

Elsewhere we have offered a model of reflective consciousness (Bickhard 1980a, Campbell and Bickhard 1986). There is no need to recapitulate that model here. For our current purposes, all that is needed is to recognize that reflective consciousness exists. And it is impossible to deny that there is such a thing: any reflection on the nature of consciousness already settles the issue. Here, what matters is the role that reflective consciousness has to play in the development of rationality.

3.3. *Critical principles and rationality*

The interactivist model of rationality is based on the recognition that a metarecursive system tends to construct internal surrogates for selection pressures. Not all of the constructions generated in a variation and selection process are going to be acceptable. Yet checking them out in the environment can be costly and risky. The system would be better off were it able to develop knowledge about the sorts of constructions that constitute errors. Then it could internalize the processes of variation and selection; it could try out new constructions internally, against its knowledge of possible errors. A metarecursive system can recognize potential errors and use its knowledge of possible errors to modify its own construction process.

When we can articulate such knowledge of error, and use it to evaluate verbally stated proposals, we say that it gives us grounds for criticism. But knowledge of potential error does not have to be articulable in order to be brought to bear against candidate knowledge or possible constructions. We will use the term *critical principles* to cover any sort of knowledge of types of errors (Bickhard 1991a, forthcoming). Critical principles are not positive knowledge—knowledge of what works or what is true. They are negative knowledge—knowledge of what to avoid.

Critical principles come in all shapes and sizes. Some are domain-general, like principles of inferential validity in logic and mathematics, or principles of logical necessity. Others are quite domain-specific: avoiding laws that are vague or unconstitutional; staying away from various kinds of engineering failures in the design of machines; preventing unmaintainable organization in computer programs; and so on. In the study of expertise we are concerned with what Feldman (1980) calls *nonuniversal* skills—those not developed by all normal human beings; those that normally have to be taught. Any account of expertise will have to contend with the relatively domain-specific skills, those that have to do with fire fighting or computer programming or jazz improvising or ferreting out counterfeit 19th century British stamps.

3.4. *Hierarchies of critical principles*

Knowledge of critical principles is fallible, just like any other knowledge. Critical knowledge must be constructed, just like any other knowledge. And critical knowledge is subject to reflection, just like any other knowledge. Reflection on critical principles yields higher-order critical principles, or meta-critical principles, that are *about* other critical principles. Higher-order critical principles are negative knowledge about critical principles themselves, knowledge of the ways other critical principles might be in error.

Meta-critical principles can support or undermine the critical principles to which they apply. Sometimes they do both—for instance, a higher-order principle could affirm a lower-order critical principle at a deeper level of analysis, while restricting the scope of that lower-order principle.

The principle of symmetry, a critical principle that applies to theories in fundamental physics, has been affirmed and strengthened over the past few decades through the development of principles about desirable, more specific forms of symmetry. By contrast, the meta-critical principles that have developed in psychology over the past 50 years (such as the principle that any theory in psychology must be concerned with internal organizations, in part because any pattern of observable behaviour can be produced by an unbounded variety of different internal organizations) have entirely demolished the old critical principle, strongly promoted by the behaviourists, that a science of psychology can only be about observable patterns of behaviour.

The critical principle that in any acceptable biological theory, a gene must be identified with a physical stretch of DNA, has had a more complex history. Subsequent developments have affirmed the basic assumption that genes are ontologically realized in DNA. But they have also undermined the claim that functional genes lie in one-to-one correspondence with stretches of DNA. Shifts in the 'reading frame' that is used when DNA controls the construction of RNA and proteins allow different functional genes to coexist in the same physical stretch of DNA. For instance, many genes in the bacterium *E. coli* actually overlap their neighbours by four base pairs (Eyre-Walker 1995). The early principle that assumed a one-to-one mapping between genes and sequences of DNA base pairs has been affirmed in one respect and refuted in another, by subsequent principles that require a more complex relationship between functional genes and physical genes.

3.5. *Negative knowledge versus positive knowledge*

Negative knowledge doesn't behave the same way that positive knowledge does. It can be rational to apply critical principles in a rigorous fashion—even when it is not rational to believe that anything fully satisfies those principles. A number of basic questions in the history and philosophy of science illustrate this point.

3.5.1. *Realism in science.* Should scientific theories be taken to make ontological claims about the way things really are, or should they be understood merely as convenient instruments for making empirical predictions? Scientific realism is a highly contentious issue in contemporary philosophy (Fine 1984, Harré 1970, Hollis and Lukes 1982, Laudan 1977, 1984, Leplin 1986). Traditionally, realists have claimed that a successful theory, one that is consistent with the available data and able to offer plausible explanations of phenomena within its scope, must also correctly describe the unobservable aspects of the world about which it makes claims. Yet many a successful theory has been overturned later on. Just in physics, the long roll of the deposed would have to include the Aristotelian doctrine of natural place, Newtonian mechanics, Newtonian gravity, phlogiston models of fire, caloric accounts of heat, and theories of the luminiferous ether. What makes us think that today's established theories will escape the same fate? And if current theories are also destined to be overturned, what rational grounds do we have for believing in the reality of the entities that our theories propose? Is it rational, for instance, to believe that quarks exist?

Some have concluded from such historical considerations that we have no rational grounds for believing in quarks—any more than we previously had for believing in phlogiston, or the luminiferous ether. And it might seem only a mild stretch to the further conclusion that neither realism nor truth plays any rational role in science (Laudan 1977, 1984, van Fraassen 1980). If we cannot have rational grounds for believing that the entities posited by our current theory are real, or for believing that our current theory is true, why bother with truth or realism at all?

3.5.2. *Critical principles in scientific rationality.* Our conception of negative knowledge suggests an alternative. It can be useful to discover in what ways a model fails to satisfy critical principles of realism or truth, even if we can never know whether our model correctly describes unobservable realities. It can be useful to know in what ways a theory fails to satisfy criteria of realism, or in what ways it survives exposure to these surrogate selection pressures—for the time being, at least.

Initially the theory of quarks was a bookkeeping device for sorting observed and unobserved particle interactions. It was purely instrumental, serving to compute the right answers. Although its originators had hopes, quark theory was not widely accepted as a realistic account of anything. In fact, there was a rival conception at the time (Reggeon theory) that was completely equivalent to quark theory from an instrumental standpoint. But when quark theory was taken seriously from an ontological point of view, it was found to have certain high-energy consequences that did differentiate it from Reggeon theory. When experiments were performed to test those consequences, the results were consistent with tentative realism about quarks, not tentative realism about Reggeons. So treating quark theory as making ontologically realistic claims led to a major advance in physics (Dodd 1984, Riordan 1992). Yet successfully withstanding those experimental tests does not show with certitude that quark theory is correct, so it gives us no certainty that quarks truly exist.

Our point can be put more generally. Critical principles, when stated as propositions, can be uncomputable (Boolos and Jeffrey 1989, Cutland 1980). There may be no finite series of computational steps sufficient to establish that the critical principle is satisfied for a given case. Tests of the high-energy consequences of quark theories did not prove or verify the existence of quarks. But quark theory might have failed these tests for the principle of realism; in that case, it would have been rational to conclude that quarks do not exist. Critical principles give us a basis for rejecting certain theories, even when it is impossible to prove that a theory satisfies those principles.

Some critical principles are partially computable; others may be wholly uncomputable. Popper (1959) pointed out that scientific theories contain universally quantified statements—of the form 'All Xs are Ys'—and that universally quantified statements can be falsified by empirical data, but not verified. Scientific critical principles are meta-hypotheses about what counts as a good theory. Theories that fail a critical principle of realism are refuted or falsified in that regard. Yet it can never be ascertained that a theory has been verified with regard to realism.

A viable theory of rationality in general will have to include negative knowledge—in the form of critical principles. So, more specifically, will accounts of scientific rationality, and of expertise. A domain of knowledge, properly understood, has distinctive heuristics for generating new hypotheses and selection criteria for evaluating them (Campbell and Bickhard 1992, Campbell 1993, Keil 1990). That is, any domain of knowledge includes critical principles. From this standpoint, rational knowledge simply means knowledge that avoids known sorts of error by satisfying known critical principles. Rational *positive* knowledge isn't knowledge of truth, or knowledge that is certain, or even knowledge that is probably true in some sense. It is knowledge that avoids known errors. Rational positive knowledge is knowledge that fits within a framework of negative knowledge—within a hierarchy of critical principles.

3.5.3. *Progress and cumulation.* Most observers believe that scientific progress has taken place, over the long haul at least. Some developmental psychologists have gone so far as to insist that changes in knowledge or skill do not constitute development unless they are progressive. But what does progress consist of? It can't be defined as movement toward an ideal endpoint, because that would require the developing system (or the system that is assessing that developing system) to know that endpoint in advance. And we don't know in advance what the truth is, or where it might be located. Consequently, it won't do to characterize progress as coming nearer to the truth. Attempts to characterize progress or rationality in terms of closer approaches toward truth, or more precise approximations to truth, have failed (Bartley 1987, Elster 1981, Hooker 1995, Laudan 1984, Suppe 1977).

Still, the construction of knowledge is progressive—in a negative sense. Culturally or individually, progress means coming to avoid more kinds of errors. There is a good deal more cumulation in the reflexive applicability of hierarchies of critical principles than there is cumulation in positive knowledge. Positive knowledge is often overthrown as new sorts of error are discovered; negative knowledge is more likely to survive (though its survival is by no means certain).

Aristotle's physics incorporated laws that varied from one physical place to another. Within the sublunary sphere, things could come into being and pass away; within the heavenly spheres, they could not. Newton showed that invariance with regard to location was an important error criterion. Aristotle's physics failed this critical principle; Newton's survived it. Newton unified conceptions of earth and the heavens by showing that common laws governed both: the same kinds of physical laws explained apples falling and planets circling the sun.

Einstein's Special Theory of Relativity retained the Newtonian principle of invariance over location, but added to it a principle of invariance for velocity (which is the first time derivative of location). Newton's mechanics failed this criterion; the Special Theory of Relativity survived it. Einstein's General Theory of Relativity added another critical principle: invariance over acceleration. The Special Theory doesn't survive the application of this criterion, but the General Theory does (in fact, it displays invariance over all of the higher-order time derivatives of location). In the progression from Aristotle's physics to the General Theory of Relativity, it is the critical principles that survive and cumulate, not the positive knowledge contained and expressed in the various theories. Each theory was overthrown in favour of a successor that satisfied a new critical principle in addition to those previously accepted.

3.5.4. *Negative knowledge for the individual.* As it goes for the history of science, we argue, so it goes for the development of the individual. Again, negative knowledge is most important in the long run. Positive knowledge often gets replaced because it fails the tests imposed by new negative knowledge. Meanwhile, the negative knowledge persists: what counts as error changes much less often than what avoids error. When there are changes in negative knowledge, they tend to add to it rather than overturn old conceptions of error. Negative knowledge is the skeleton around which positive knowledge is organized and the base on which further development occurs.

From our standpoint, expertise largely consists in a vast knowledge of sorts of errors that can occur and that can be made in myriads of specific circumstances within the domain. Non-experts will tend to blunder into such specific, perhaps rare, errors. All too often, that is how experts acquired their knowledge as well. If we want to understand how people become experts, we need to keep negative knowledge and its development at the forefront.

4. Expertise

We need to know, then, how experts make rational choices and how they generalize in their domain; what's more, we need to know how they acquire such abilities and continue to improve at them. Rationality, we have contended, has as its skeleton a hierarchy of critical principles; it consists of negative knowledge about possible errors to avoid. Generalization means generating appropriate problem solutions for new situations based on their similarity to old solutions previously used in old situations. The ability to generalize (and to learn to generalize) imposes its own constraints on system architecture, constraints that we sketched earlier (Bickhard and Campbell in press).

A crucial function of negative knowledge in expertise is to constrain generalization. Generalization involves heuristics for knowledge construction, and, like any heuristics, these are inherently fallible. Generalization isn't rational unless knowledge of possible errors is involved. Critical principles allow potential generalizations to be tried out against knowledge of errors before they have to be tried out in the real world.

4.1. *Teaching and designing*

4.1.1. *Negative knowledge in training.* Most forms of expertise require extensive teaching or apprenticeship, although neither guarantees attainment of the highest levels. Today negative knowledge rates little attention from teachers or trainers. Positive knowledge gets the spotlight, as might be expected in a culture like ours that is permeated with empiricism. And from the empiricist standpoint, knowledge is something to be inserted, impressed, or poured into empty minds (Popper 1972). The empiricist grip remains so strong that a recent, highly sophisticated discussion of the effects of deliberate practice in the attainment of expertise can cite traditional lore from baseball coaches and violin teachers about the importance of identifying and working on one's weaknesses, yet insist on the correctness of a model of skill acquisition as the smooth accumulation of positive knowledge over time (Ericsson *et al.* 1993).

Such attitudes notwithstanding, teaching about misconceptions by presenting countervailing examples and principles can be most useful. For instance, students' misconceptions in science are a prevalent, persistent, and troubling phenomenon in education (Berliner and Casanova 1987, Brumby 1984, Champagne *et al.* 1983, Confrey 1990, Novak 1987, Perkins and Simmons 1988). Such misconceptions frequently maintain their hold on students, even after they have thoroughly mastered the standard positive-knowledge curriculum. Explicitly teaching negative knowledge helps students to overthrow such misconceptions—for instance, the implicit Lamarckian assumptions that they tend hold alongside their 'textbook' evolutionary theory (Wu 1993). It enables students not just to learn what the accepted positive knowledge is, but also to learn the principles in terms of which the misconceptions are in error, and, consequently, in terms of which the accepted knowledge is a better alternative. It gives students good reasons for rejecting misconceptions which under other circumstances they might never have realized they held.

But the value of teaching critical principles is not restricted to issues on which one's current students hold misconceptions. In a deep sense we understand something only to the extent that we know why prima facie alternative conceptions ought to be rejected. Otherwise all we have is rote knowledge; we know positive answers and procedures without knowing what makes them correct, or successful, or better than other possibilities. We don't know what errors they avoid, or how they avoid those errors better than other answers would. It can be useful, then, to teach relevant critical principles even if no particular student currently harbours the misconceptions they guard against. For instance, it is useful to teach students the inadequacy of any psychological theory that avoids specifying inner mechanisms—even if none of them subscribes to behaviourism. Knowing the critical principles gives students a base and a framework for developing new knowledge through variation and selection. And the ability to keep developing is a crucial mark of genuine expertise (Bereiter and Scardamalia 1993).

Imagine what would happen if we sought to eliminate negative knowledge from training altogether, instead of making it peripheral to training or merely implicit in our practices, as it usually is now. We could try to teach expertise while strictly adhering to the principles of 'instructional design' (Gagné *et al.* 1992, Leshin *et al.* 1992, Polson 1987). According to these precepts, we should take expert skill in the domain of interest and break it down into numerous units or modules of positive knowledge. The modules would have a hierarchical organization, and consequently there would be a proper sequence in which they should be taught, so that all prerequisites would be in place before moving on to a skill that requires them. Each module would be practised to completely correct performance before moving on to the next. In the instructional design worldview, error is supremely uninteresting. Errors will occur, and they must be corrected on the spot, for reasons of 'reinforcement' (Gagné *et al.* 1992). But the types of errors students make, and ways to avoid such errors, play no role in instructional design.

From an interactivist standpoint, an instructional-design model for training experts would be entirely wrongheaded. We don't learn how to hit a baseball by practising each component head, shoulder, arm, wrist, hip, and leg motion to an error-free criterion, then combining all of these components. If knowledge is a kind of action, albeit internal action (taking the form of flows of information and control within a system), then the utter unsuitability of practising all components to perfection, then synthesizing, will hold for knowledge in general.

Moreover, we would argue that the errors human beings make in mastering a domain are not all avoidable imperfections, slips, or 'performance limitations'. Instead, the *best possible* developmental pathway towards expertise necessarily incorporates errors. Were extremely clever and foresighted trainers to design a training procedure for some kind of expert skill that elicited error-free performance at every step, and denied the trainee any opportunity to make mistakes, the results would be self-defeating. The trainee, deprived of the opportunity to make errors, would have a difficult time coming up with critical principles that identify the kinds of errors to avoid. He or she would never develop any sort of rationality about the domain in question.

4.1.2. *Negative knowledge in expert systems.* Just as negative knowledge is usually overlooked when human expertise is being analysed, it has rarely figured in the design of expert systems (although adaptability is a significant theme in the expert systems projects presented in this issue, negative knowledge is not). What's usually prized is more accurate and more extensive bases of positive knowledge. Although frozen accumulations of positive knowledge can be useful for some applications, there is no way they will ever capture the general ability of experts to become more expert.

Indeed, it has been noted that knowledge bases generally don't scale up; in other words, they don't form a useful base for developing and incorporating new knowledge. Whatever their utility for some synchronic task domain, they can't handle the diachronic task of developing more expert knowledge. Taken to its logical extreme, the drive for positive knowledge leads to the utopian project of building a universal knowledge base that could be adequate for every conceivable purpose (Lenat and Guha 1988, Lenat *et al.* 1988, Lenat and Feigenbaum 1991; for an evaluation of this project, see Bickhard and Terveen 1995).

A positive-knowledge strategy may produce workable expert systems, if the domain of expertise is narrow enough and the needed knowledge can be successfully rendered into encoded rules. That is, it may work if the goal is to provide advice on obeying corporate capital accounting rules or to specify how the load should be placed inside a C-5 military cargo jet (although even in tractable cases like these, unanticipated confusion can still arise for the user—see Whitley 1996). But models of human expertise have to explain how rational generalization develops, as well as how it happens. More ambitious expert systems will need to manifest this capacity and be able to acquire more of it. No progress will be made in these endeavors without addressing issues of negative knowledge.

4.2. *Minsky's negative metaknowledge*

Our ideas may strike some AI researchers as similar to Minsky's (1983, 1986) proposals about negative metaknowledge. We find the resemblance only slight. Minsky's inspiration was not evolutionary epistemology, but Freud's (1950) conception of a 'censor' that removes overt expressions of dangerous primitive impulses from dreams, or finds an 'acceptable' disguise for them. We agree with Minsky that it is desirable to know what to avoid, but the area of agreement abruptly terminates thereafter. Even if we disregard the other troubles that afflict a Freudian cognitive architecture, Freudian censorship is based on threats of punishment, or of being overwhelmed by instinctual energies. It is obvious how Freudian censorship would foster learning to avoid what is dangerous. It is not so easy to see how it would encourage the development of critical principles that identify broad categories of errors, and even explain what is erroneous about them.

By contrast, evolutionary epistemology, of the sort we have previously outlined, is able to cope with issues of conceptual or scientific or logical error. And our interactivist conception includes a hierarchy of knowing levels—levels of reflective consciousness—which makes possible a hierarchy of critical principles, or negative knowledge *about* negative knowledge. A hierarchy of critical principles makes it easy to accommodate the developmental construction and cumulation of critical principles. Yet there is nothing like knowing levels in a Freudian cognitive architecture, or in Minsky's society of mind. Nor does a conception of negative knowledge as censorship have any place for considerations about the computability of critical principles. Our account affords a much more highly developed treatment of negative knowledge than Minsky's does.

5. Conclusion

To understand expertise in human beings, or capture its attributes in expert systems, we need to understand rationality and generalization. Conventional AI and cognitive science are having a hard enough time modelling rationality and generalization synchronically, as end states. Yet a full understanding of expertise requires models of the diachronic aspects of rationality and generalization—how they develop. If our analysis is correct, the diachronic difficulties are insuperable within standard approaches. There has to be another way.

An alternative conception of generalization is a major undertaking, not to be attempted within the confines of this article. We have, however, laid out the unsolved problems that generalization poses for standard approaches. Standard conceptions can't explain the emergence of new representations, or of new topologies in representational spaces, yet the development of generalization requires both kinds of novelty.

We have proposed a treatment of rationality that arises straightforwardly out of the interactive account of knowledge. Knowledge is basically 'knowing how', and is constituted as system organization. Because interactive knowledge emerges out of underlying system organization, the diachronic troubles that bedevil standard approaches do not arise. Mental representation can emerge out of system organization that isn't already representational. New knowledge is possible, and so are genuine development and learning. There are none of the dead ends that encoding-based approaches run into sooner or later.

From interactivism it directly follows that new knowledge can't be impressed on the knowing system by the environment; some form of evolutionary constructivism must be correct. And negative knowledge is necessary to guide the constructive process. Negative knowledge, in the form of a hierarchy of critical principles, is the core of rationality, and rationality is integral to the development of expertise. Although negative knowledge is a necessary aspect of expertise, it is largely neglected by psychologists and designers. It needs a good deal more attention, whether our focus is on the development of human expertise, or on training that fosters the development of expertise, or on the design of expert systems.

Acknowledgments

Thanks to Steve Krawiec for information about the 'reading frame' for genes.

References

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

Aristotle. (1941) De anima [On the soul]. In R. McKeon (ed.) *The Basic Works of Aristotle* (New York: Random House) pp. 533–603. (Originally published c. 325 BC.)

Baars, B. J. (1986) *The Cognitive Revolution in Psychology* (New York: Guilford Press).

Bartley, W. W. III (1987) Theories of rationality. In G. Radnitzky and W. W. Bartley III (eds) *Evolutionary Epistemology, Theory of Rationality, and the Sociology of Knowledge* (La Salle, IL: Open Court) pp. 205–214.

Bereiter, C. and Scardamalia, M. (1993) *Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise* (Chicago: Open Court).

Berliner, D. and Casanova, U. (1987) How do we tackle kids' science misconceptions? *Instructor*, November/December: 14–15.

Bickhard, M. H. (1979) On necessary and specific capabilities in evolution and development. *Human Development*, **22**: 217–224.

Bickhard, M. H. (1980a) A model of developmental and psychological processes. *Genetic Psychology Monographs*, **102**: 61–116.

Bickhard, M. H. (1980b) *Cognition, Convention, and Communication.* (New York: Praeger).

Bickhard, M. H. (1991a) A pre-logical model of rationality. In L. Steffe (ed.) *Epistemological Foundations of Mathematical Experience* (New York: Springer). pp. 68–77.

Bickhard, M. H. (1991b) The import of Fodor's anti-constructivist argument. In L. Steffe (ed.) *Epistemological Foundations of Mathematical Experience* (New York: Springer) pp. 14–25.

Bickhard, M. H. (1992) How does the environment affect the person? In L. T. Winegar and J. Valsiner (eds) *Children's Development within Social Context: Metatheory and Theory* (Hillsdale, NJ: Erlbaum) pp. 63–92.

Bickhard, M. H. (1993) Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, **5**: 285–333.

Bickhard, M. H. (forthcoming) Critical principles: On the negative side of rationality. In W. Herfel and C. A. Hooker (eds) *Beyond Ruling Reason: Non-formal Approaches to Rationality.*

Bickhard, M. H. and Campbell, R. L. (in press) Topologies of learning and development. New Ideas in Psychology.

Bickhard, M. H. and Terveen, L. (1995) *Foundational issues in Artificial Intelligence and Cognitive Science—Impasse and Solution* (Amsterdam: North-Holland)

Bochenski, I. M. (1970) *A history of formal logic* (New York: Chelsea).

Boolos, G. S. and Jeffrey, R. C. (1989) *Computability and logic*, 3rd edn (Cambridge: Cambridge University Press).

Braine, M. D. S. and Rumain, B. (1983) Logical reasoning. In J. H. Flavell and E. M. Markman (eds) *Handbook of Child Psychology, Vol. III: Cognitive Development* (New York: Wiley) pp. 263–340.

Bringsjord, S. and Bringsjord, E. (1996) The case against AI from imagistic expertise. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 383–397.

Brown, H. I. (1988) *Rationality* (London: Routledge).

Brumby, M. N. (1984) Misconceptions about the concept of natural selection by medical biology students. *Science Education*, **68**: 492–503.

Campbell, D. T. (1974) Evolutionary epistemology. In P. A. Schilpp (ed.) *The Philosophy of Karl Popper* (LaSalle, IL: Open Court) pp. 413–463.

Campbell, R. L. (1991) Does class inclusion have mathematical prerequisites? *Cognitive Development*, **6**: 169–194.

Campbell, R. L. (1993) Epistemological problems for neo-Piagetians. In A. Demetriou, A. Efklides and M.

Platsidou, *The Architecture and Dynamics of Developing Mind: Experiential Structuralism as a Frame for Unifying Cognitive Developmental Theories. Monographs of the Society for Research in Child Development*, 58 (5–6, ser. no. 234), pp. 168–191.

Campbell, R. L. and Bickhard, M. H. (1986) *Knowing Levels and Developmental Stages* (Basel, Switzerland: Karger).

Campbell, R. L. and Bickhard, M. H. (1987) A deconstruction of Fodor's anticonstructivism. *Human Development*, **30**: 48–59.

Campbell, R. L. and Bickhard, M. H. (1992) Types of constraints on development: An interactivist approach. *Developmental Review*, **12**: 211–238.

Campbell, R. L. and Di Bello, L. A. (1996) Studying human expertise: Beyond the binary paradigm. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 277–291.

Champagne, A. B., Gunstone, R. F. and Klopfer, L. E. (1983) Naive knowledge and science learning. *Research in Science and Technological Education*, **1**: 173–183.

Cherniak, C. (1986) *Minimal Rationality* (Cambridge, MA: MIT Press).

Clark, A. (1993) *Associative Engines: Connectionism, Concepts, and Representational Change* (Cambridge, MA: MIT Press).

Confrey, J. (1990) A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education*, **16**: 3–56.

Cutland, N. J. (1980) *Computability* (Cambridge: Cambridge University Press).

Dancy, J. (1985) *Contemporary Epistemology* (New York: Basil Blackwell).

Dennett, D. C. (1991) *Consciousness Explained* (Boston, MA: Little Brown).

Dodd, J. E. (1984) *The Ideas of Particle Physics* (Cambridge: Cambridge University Press).

Dreyfus, H. L. (1967) Why computers must have bodies in order to be intelligent. *Review of Metaphysics*, **21**: 13–32.

Dreyfus, H. L. (1982) Introduction. In H. L. Dreyfus (ed.) *Husserl: Intentionality and Cognitive Science* (Cambridge, MA: MIT Press) pp. 1–27.

Dreyfus, H. L. (1991) *Being-in-the-World: A Commentary on Heidegger's* Being and time, Division I (Cambridge, MA: MIT Press).

Dreyfus, H. L. and Dreyfus, S. E. (1986) *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (New York: The Free Press).

Elster, J. (1981) Rationality. In G. Fløistad (ed.) *Contemporary Philosophy*, Vol. 2 (Dordrecht: Martinus Nijhoff).

Ericsson, K. A., Krampe, R. T. and Tesch-Römer, C. (1993) The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, **100**: 363–406.

Eyre-Walker, A. (1995) The distance between *Escherichia coli* genes is related to gene expression levels. *Journal of Bacteriology*, **177**: 5368–5369.

Feldman, D. H. (1980) *Beyond Universals in Cognitive Development* (Norwood, NJ: Ablex).

Fine, A. (1984) The natural ontological attitude. In J. Leplin (ed.) *Scientific Realism* (Berkeley: University of California Press) pp. 83–107.

Fodor, J. (1972) Some reflections on L. S. Vygotsky's *Thought and Language. Cognition*, **1**: 83–95.

Fodor, J. (1975) *The language of thought* (New York: Crowell).

Fodor, J. (1981) The present status of the innateness controversy. In J. Fodor (ed.) *RePresentations* (Cambridge, MA: MIT Press) pp. 257–316.

Fodor, J. (1987) *Psychosemantics* (Cambridge, MA: MIT Press).

Fodor, J. (1990a) *A Theory of Content* (Cambridge, MA: MIT Press).

Fodor, J. (1990b) Information and representation. In P. P. Hanson (ed.) *Information, Language, and Cognition* (Vancouver: University of British Columbia Press) pp. 175–190.

Fodor, J. and Pylyshyn, Z. (1981) How direct is visual perception? Some reflections on Gibson's ecological approach. *Cognition*, **9**: 139–196.

Ford, K. M. and Hayes, P. J. (eds) (1991) *Reasoning Agents in a Dynamic World: The Frame Problem* (Greenwich, CT: JAI Press).

Ford, K. M. and Pylyshyn, Z. (eds) (in press) *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence* (Norwood, NJ: Ablex).

Freud, S. (1950) *The Interpretation of Dreams* (New York: Modern Library) (Original published 1900.)

Gagné, R. M., Briggs, L. J. and Wager, W. W. (1992) *Principles of Instructional Design*, 4th edn (Fort Worth, TX: Harcourt Brace Jovanovich).

Gentner, D. and Grudin, J. (1985) The evolution of mental metaphors in psychology. *American Psychologist*, **40**(2): 181–192.

Gentner, D. and Jeziorski, M. (1994) The shift from metaphor to analogy in Western science. In A. Ortony (ed.) *Metaphor and thought*, 2nd edn. (Cambridge: Cambridge University Press) pp. 447–480.

Gentner, D. and Markman, A. B. (1995) Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (ed.) *Similarity in Language, Thought and Perception* (Milan: Brepols).

Gentner, D. and Rattermann, M. J. (1991) Language and the career of similarity. In S. A. Gelman and J. P. Byrnes (eds) *Perspectives on Language and Thought: Interrelations in Development* (Cambridge: Cambridge University Press) pp. 225–277.

Geroch, R. (1985) *Mathematical physics* (Chicago: University of Chicago Press).

Hahlweg, K. and Hooker, C. A. (Eds.) (1989). *Issues in Evolutionary Epistemology* (Albany: State University of New York Press).

Hammond, K. (1989) *Case-based Planning: Viewing Planning as a Memory Task* (New York: Academic Press).

Harré, R. (1970) *The Principles of Scientific Thinking* (Chicago: University of Chicago Press).

Hewett, R. (1996) Using programming expertise for controlling software synthesis. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 293–318.

Hocking, J. G. and Young, G. S. (1961) *Topology* (Reading, MA: Addison-Wesley).

Hollis, M. and Lukes, S. (1982) *Rationality and Relativism* (Cambridge, MA: MIT Press).

Hooker, C. A. (1995) *Reason, Regulation, and Realism: Towards a Regulatory Systems Theory of Reason and Evolutionary Epistemology* (Albany, NY: SUNY Press).

Hume, D. (1888) *A Treatise of Human Nature* L. A. Selby-Bigge, ed. (Oxford: Clarendon Press) (Originally published 1739.)

James, I. M. (1987) *Topological and Uniform Spaces* (New York: Springer-Verlag).

Johnson-Laird, P. N. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge, MA: Harvard University Press).

Keil, F. C. (1990) Constraints on constraints: Surveying the epigenetic landscape. *Cognitive Science*, **14**: 135–168.

Kneale, W. and Kneale, M. (1986) *The Development of Logic* (Oxford: Clarendon).

Kolodner, J. L. and Simpson, R. L. (1989) The MEDIATOR: Analysis of an early case-based problem solver. *Cognitive Science*, **13**: 507–549.

Laudan, L. (1977) *Progress and its Problems* (Berkeley: University of California Press).

Laudan, L. (1984) A confutation of convergent realism. In J. Leplin (ed.) *Scientific Realism* (Berkeley: University of California Press) pp. 218–249.

Leake, D. B. (1996) Experience, introspection, and expertise: Learning to refine the case-based reasoning process. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 319–339.

Lenat, D. B. and Feigenbaum, E. A. (1991) On the thresholds of knowledge. *Artificial Intelligence*, **47**: 185–220.

Lenat, D. B. and Guha, R. (1988) *The world according to CYC*. Technical Report No. ACA-AI-300–88. Austin, TX: MCC.

Lenat, D. B., Guha, R. and Wallace, D. (1988) *The CycL representation language*. Technical Report No. ACA-AI-302–88. Austin, TX: MCC.

Leplin, J. (1986) Methodological realism and scientific rationality. *Philosophy of Science*, **53**: 31–51.

Leshin, C. B., Pollock, J. and Reigeluth, C. M. (1992) *Instructional Design Strategies and Tactics* (Englewood Cliffs, NJ: Educational Technology Publications).

Loewer, B. and Rey, G. (eds) (1991) *Meaning in Mind: Fodor and his Critics* (Oxford: Basil Blackwell).

Mackie, J. L. (1985) Rationalism and empiricism. In J. L. Mackie (ed.) *Logic and Knowledge* (London: Oxford University Press).

Mathews, R. C., Lane, I. M., Roussel, L. G., Nagy, M. S., Haptonstahl, D. E. and Brock, D. B. (1996) Using conscious reflection, group processes, and AI to facilitate the development of expertise. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 259–276.

Medin, D. L., Goldstone, R. L. and Gentner, D. (1993) Respects for similarity. *Psychological Review*, **100**: 254–278.

Medin, D. L. and Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review*, **85**: 207–238.

Minsky, M. (1983) Jokes and the logic of the cognitive unconscious. In R. Groner, M. Groner and W. F. Bischoff (eds) *Methods of Heuristics* (Hillsdale, NJ: Erlbaum) pp. 171–193.

Minsky, M. (1986) *The Society of Mind* (New York: Simon and Schuster).

Moser, P. K. (1987) *A Priori Knowledge* (New York: Oxford University Press).

Novak, J. (ed.) (1987) *In The Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics*, Ithaca, NY: Department of Education, Cornell University.

Perkins, D. N. and Simmons, R. (1988) Patterns of misunderstanding: An integrative model for science, math, and programming. *Review of Educational Research*, **58**: 303–326.

Polson, P. G. (1987) A quantitative theory of human-computer interaction. In J. M. Carroll (ed.) *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction* (Cambridge, MA: MIT Press) pp. 184–235.

Popper, K. R. (1959) *The Logic of Scientific Discovery* (New York: Harper and Row).

Popper, K. R. (1972) The bucket and the searchlight: Two theories of knowledge. In K. Popper (ed.) *Objective Knowledge: An Evolutionary Approach* (Oxford: Clarendon) pp. 341–361.

Prietula, M. J., Vicinanza, S. S. and Mukhopadhyay, T. (1996). Software-effort estimation with a case-based reasoning process. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 341–363.

Pylyshyn, Z. W. (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Norwood, NJ: Ablex).

Rich, E. and Knight, K. (1991) *Artificial Intelligence* (New York: McGraw-Hill).

Riesbeck, C. K. and Schank, R. C. (1989) *Inside Case-based Reasoning* (Hillsdale, NJ: Erlbaum).

Riordan, M. (1992) The discovery of quarks. *Science*, **256**: 1287–1293.

Shalin, V. L. and Bertram, D. A. (1996) Functions of expertise in a medical intensive care unit. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 209–227.

Shanon, B. (1988) On the similarity of features. *New Ideas in Psychology*, **6**, 307–321.

Suppe, F. (1977) *The Structure of Scientific Theories*, 2nd edn (Urbana: University of Illinois Press).

Terwilliger, R. F. (1968) *Meaning and Mind: A Study in the Psychology of Language* (New York: Oxford University Press).

Tversky, A. (1977) Features of similarity. *Psychological Review*, **84**: 327–352.

van Fraassen, B. C. (1980) *The Scientific Image* (Oxford: Clarendon).

Whitley, E. A. (1996) Confusion, social knowledge, and the design of intelligent machines. *Journal of Experimental and Theoretical Artificial Intelligence*, **8**: 365–381.

Wu, P. (1993) *The Rationality Model and Students' Misconceptions*. Unpublished doctoral dissertation, Department of Educational Psychology, University of Texas at Austin.

Wuketits, F. M. (1990) *Evolutionary Epistemology and its Implications* (Albany: State University of New York Press).