Birkbeck
UNIVERSITY OF LONDON

BIROn - Birkbeck Institutional Research Online

Cooper, Richard P. (2007) The role of falsification in the development of cognitive architectures: insights from a Lakatosian analysis. Cognitive Science 31 (3), pp. 509-533. ISSN 0364-0213.

Downloaded from: https://eprints.bbk.ac.uk/id/eprint/6777/

Usage Guidelines:
Please refer to usage guidelines at https://eprints.bbk.ac.uk/policies.html or alternatively contact lib-eprints@bbk.ac.uk.

To appear in *Cognitive Science*, 2007

The Role of Falsification in the Development of Cognitive Architectures: Insights from a Lakatosian Analysis

Richard P. Cooper
School of Psychology
Birkbeck, University of London

Abstract: It has been suggested that the enterprise of developing mechanistic theories of the human cognitive architecture is flawed because the theories produced are not directly falsifiable. Newell attempted to sidestep this criticism by arguing for a Lakatosian model of scientific progress in which cognitive architectures should be understood as theories that develop over time. However, Newell's own candidate cognitive architecture adhered only loosely to Lakatosian principles. This paper reconsiders the role of falsification and the potential utility of Lakatosian principles in the development of cognitive architectures. It is argued that a lack of direct falsifiability need not undermine the scientific development of a cognitive architecture if broadly Lakatosian principles are adopted. Moreover, it is demonstrated that the Lakatosian concepts of positive and negative heuristics for theory development and of general heuristic power offer methods for guiding the development of an architecture and for evaluating the contribution and potential of an architecture's research program.

1. Introduction

1.1 Newell's call for cumulative research programs

In a now famous commentary reflecting on a set of papers presented at a symposium on information processing psychology, Newell (1973) expressed serious concern at the direction in which theorizing within cognitive psychology was progressing. He argued that impressive empirical work was leading to a detailed understanding of the processes underlying specific cognitive phenomena, but that this understanding was not being integrated into wider cognitive theory. The knowledge associated with each phenomenon was isolated and ultimately fragmentary. Newell argued that the then contemporary methods would not yield progress on (what he regarded as) the important question, and that such progress required the development of integrative theories that spanned cognitive domains. Newell's argument evolved into a call for Unified Theories of Cognition (UTCs: Newell, 1990) – theories that specify "a single set of mechanisms for all of cognitive behavior" (Newell, 1990, p. 15) – and the assertion that UTCs would take the form of mechanistic cognitive architectures. To clarify and support his arguments Newell also presented the Soar cognitive architecture as a candidate UTC.

The shift from single domain theories to UTCs involves more than a shift in theoretical focus. It involves shifts in research methodology and criteria for measuring scientific progress. Theoretical development of single domain theories within cognitive psychology can be characterized as resulting from joint processes of empirical confirmation and Popperian falsification (Popper, 1935), with theories making predictions and empirical work supporting or refuting those theories. Newell's claim (as is clear from the title of his 1973 paper: "You can't play twenty questions with nature and win") was that the development of UTCs would require an alternative methodology. Newell (1990) made this explicit by arguing that Popperian

falsification is inappropriate for the development of UTCs, and that a more adequate model for theory development is that of a cumulative research program, as described by Lakatos (1970).[1]

1.2 The need for a systematic methodology

The move from a Popperian to a Lakatosian model of scientific progress is highly significant. Newell's rationale was that theories cumulate: "They are refined and reformulated, corrected and expanded" (Newell, 1990, p. 14). Certainly examples of theory refinement and expansion abound in Lakatos' historical analyses of scientific progress (see Lakatos, 1970), but Lakatos' primary concern was with the differentiation of science from non-science, and specifically with whether Popperian falsification was a necessary feature of a scientific theory. Newell (1990) argued against falsification by citing Lakatos, but he did not provide a detailed demonstration that Soar – his candidate UTC – did indeed adhere to Lakatosian principles. In fact, it is possible to give a broadly Lakatosian reconstruction of the development of Soar (see Cooper, 2006), but the issue of falsification is more real for Soar and UTCs in general than for single domain theories. Indeed, there remains reluctance within some sectors of the cognitive sciences to accept research on cognitive architectures as valid science, and at least some of this reluctance may be traced to the view that cognitive architectures are not falsifiable, and hence are more akin to belief systems than scientific theories (e.g., Hunt & Luce, 1992; Vere, 1992).

Historically, the motivation for the method of falsification arose from the Logical Positivists' attempts to differentiate the science of Einstein and Newton from what they viewed as the pseudo-science of Freud and Marx, coupled with Popper's (1935) critique of confirmation as a method for determining the truth of universally quantified statements. From this perspective, falsification is just a method and one can justify rejecting falsification provided one adopts an appropriate method in its place. Specifying such a method was precisely what Lakatos (1970) was attempting to do. Newell's argument, then, is that, at least within the field of UTCs and cognitive architectures, the methodology of Lakatos is an appropriate scientific methodology.

Methodology is an issue for the development of architectures because, as Newell (1990) himself noted, broad theories of behavior have come and gone. The behaviorist equivalent of unified theories of cognition, "grand theories" such as that of Hull (1943), were common in mainstream US psychology between 1930 and 1950, but such theories failed to survive the cognitive revolution of the 1960s. In part the demise of such grand theories can be attributed to their methodologically weak foundations, and in particular to the ease with which apparently confirmatory data for a broad theory of behavior may be selected from the thousands of known behavioral regularities while recalcitrant findings are assigned to "future research" (Cooper & Shallice, 1995). A sound methodological basis is therefore of critical importance in the development of theories of the cognitive architecture.

1.3 The structure of this paper

Lakatos' account of scientific change is not the only alternative to Popperian falsification that has been proposed (see, e.g., Laudan, 1977; Thagard, 1992). This paper's goal is not to evaluate these alternative accounts. Rather, its goals are, first, to consider the place of falsification in the development of cognitive architectures, and second, to demonstrate how Lakatosian principles can support the scientific development of cognitive architectures in the absence of falsification.

The remainder of the paper begins with a brief review of the historical background to falsification and Lakatos' conception of a scientific research program so that the role of falsification in scientific theorizing in general and cognitive architectures in particular may be more clearly evaluated. We then consider the extent to which applying the Lakatosian

[1] This is not to suggest that the model of theory development described by Lakatos and discussed throughout this paper is inappropriate for single-domain theories. Indeed, Lakatos' claim is that his model of theory development and scientific change is appropriate for characterising the development of all scientific theories.

perspective to the development of cognitive architectures might address the concerns that led Newell to originally suggest it. It is concluded first that falsification has only a secondary role in the development of cognitive architectures – the primary role belongs to prediction and empirical testing and this does not entail falsification – and second that the Lakatosian perspective has much to offer, particularly with respect to distinguishing between central assumptions and peripheral hypotheses, with its emphasis on empirical testing of peripheral hypotheses, and with its notion of a positive heuristic for theory development. We also consider several implications of the analysis for the development of cognitive architectures, concerning both methods to ensure architectural progress and the evaluation of a cognitive architecture in terms of its "heuristic power". Throughout, the arguments are illustrated with examples of architectural change taken from the development of Soar and ACT-R, two cognitive architectures whose heritage can each be traced back in the public record for more than 30 years.

2 Falsification and demarcation criteria

2.1 Falsifiability as an attribute of scientific theories

Historically, Popper, Lakatos and related philosophers were primarily concerned with specifying so-called "demarcation criteria" – criteria that might be employed to distinguish between endeavors that are generally agreed to be scientific (e.g., modern physics and chemistry) and endeavors that might claim scientific credibility, but which the majority of reputable scientists are hesitant to endorse (ranging from Freudian psychoanalytic theory and Marxist economic theory to astrology and creationism). Falsification was at the heart of Popper's account of the distinction between the two. For Popper, a scientific theory was any theory that could, in principle, be falsified.

The dominant view prior to Popper was that science was characterized by the method of induction, in which universal laws were "induced" from a set of instances. Thus, Newton might have induced the law $F = ma$ from a set of observations of objects with different masses (m) subject to different forces (F) and different resultant accelerations (a). Popper's insight was to understand the limits of induction – it is not possible to prove a universal statement (such as $F = ma$) with any number of instances – and to offer the method of falsification as a constructive alternative. According to the method of falsification, scientific theories (or laws) make predictions that may fail to match empirical observation. Thus Newtonian mechanics was scientific because one could use Newton's laws to predict how a body would behave given its mass and the forces acting upon it. These predictions could then be tested through empirical studies. Confirmation via such studies might add weight to a theory, but this is not what makes the theory scientific. What makes a theory scientific from the Popperian perspective is that its predictions might fail, in which case the theory is falsified and can be rejected.

Popper's arguments for falsification were so persuasive that falsification has effectively become enshrined in "the scientific method" as, for example, taught in many secondary school science courses, but it is necessary to distinguish between a theory that is falsifiable and falsification as a method. Consider, for example, the method of contemporary experimental psychology. Here the standard approach is to generate predictions from theory (a necessary first step in the method of falsification, and something that is only possible if a theory is indeed falsifiable), but the goal of experimental work is generally not to falsify those predictions. Rather, the goal is to confirm or obtain support for those predictions. Thus, contemporary experimental psychology appears to accept that theories must be falsifiable without adhering to the method of falsification.[2]

[2] The approach of contemporary experimental psychology is more easily related to Popper's later view – that science proceeds through the experimental testing of "bold conjectures" (Popper, 1965).

2.2 The difficulty of falsifying architectures

Popperian falsification is not appropriate for the development of cognitive architectures because, in the absence of task knowledge, architectures make few if any predictions. Thus, while architectures such as Soar or ACT-R might be used successfully to simulate performance on, for example, problem solving or reasoning tasks such as cryptarithmetic (Newell, 1990) or algebraic reasoning (Anderson, 2005), to do so requires a specification of task knowledge. Conceptually, this task knowledge amounts to a separate theory of the task being modeled, and both the task theory and the architecture in which it is embodied are required in order to make predictions. Given this, any attempt at falsification (i.e., any attempt at empirically testing the predictions of the architecture/task theory) cannot address the architecture in the absence of the task theory. If predictions are found to be false, it is unclear if the predictive failure should be attributed to the architecture or the task theory.

The situation is complicated further in the case of architectures that contain free parameters and/or implementation details to which the architectural theorist is not committed, particularly when the dominant methodological approach is to demonstrate a match between performance of the architecture and performance of human participants on equivalent tasks. If free parameters are involved, simulating behavior may amount to parameter fitting, which demonstrates merely that an architecture is consistent with the behavior being simulated (Roberts & Pashler, 2000).

Yet more complications arise from that fact that the functioning of complex systems can generally be described at several different levels (e.g., Newell, 1981; Marr, 1982; Pylyshyn, 1984). Cognitive architectures frequently span these levels. Thus Newell (1990) presented Soar as a theory at the problem space level (in which cognition involves the selection and application of operators to a problem solving state) and a theory at the symbol level (in which selection and application of operators involves the firing of condition-action rules), with the symbol level implementing the problem space level. While the use of levels of description may clarify the theoretical commitment of different aspects of an architecture, it also raises the possibility that assumptions at one level might be falsified without impacting upon those at another. We take the view that the issue of levels of description is largely orthogonal to that of falsification, and hence do not discuss this complication in any further detail.

In the limit, falsifying a cognitive architecture requires a demonstration that the architecture cannot account for some aspect of observed behavior on some task. This in turn requires a demonstration that the architecture cannot account for the behavior with all possible theories of that task, all possible parameter variations, and all possible implementations of architectural implementation details. Demonstrating that the architecture cannot account for the data with one specific theory of the task, one parameter setting or one implementation merely falsifies that specific combination of architecture, task theory, parameter setting and implementation, rather than any element of that combination. The same architecture, with a different theory of the task, different parameter setting or different implementation, may be able to account for the data. Falsification in the case of architectures, is therefore subject to the same criticism that Popper applied to the Logical Positivists' method of induction: it requires attempting to prove a universally quantified statement through a series of instances, where the universally quantified statement is that for each possible theory of a task, parameter setting and implementation, the combined architecture/task theory fails to reproduce the observed data.

2.3 Interim conclusion

Given the difficulties discussed above in falsifying architectures, it seems that we have three options. Either: 1) we accept that cognitive architectures are not scientific theories; 2) we dismiss the proposition that there exist demarcation criteria that distinguish between science and pseudo-science; or 3) we provide and justify alternative demarcation criteria that admit research on cognitive architectures as scientific.

The first option is counter-intuitive given that we have no reason to believe that the object of study (i.e., the putative information processing mechanisms of the brain) is not open to scientific investigation. If one accepts that the putative information processing mechanisms of the brain are open to scientific investigation, then one must accept that scientific theories of these mechanisms can be developed. The issue then is of finding appropriate methods to support the scientific investigation of the putative mechanisms.

Laudan (1983) has presented strong arguments in favor of the second option. His conclusion was that there are no necessary and sufficient conditions that distinguish science from pseudo-science. While this may be the case, it does not mean that we have to give up demarcation. One plausible view that is consistent with Laudan's basic position is that while there is no essential characteristic that differentiates science from pseudo-science, instances of science, like instances of Wittgenstein's games, share family resemblance (Wittgenstein, 1953). On this view, falsification may be a feature common to many scientific theories, but it is not the only feature common to many scientific theories, and more critically it is not a necessary feature of a scientific theory. Indeed, on the family resemblance view no single feature is necessary and sufficient in defining science.

It is the third option that Newell (1990) endorses with his claim that the development of cognitive architectures requires a methodology based on that proposed by Lakatos (1970).[3] Note though that in attempting to provide alternative demarcation criteria, it is not constructive to simply provide criteria that distinguish putative science from putative pseudo-science. Rather, if the criteria are to play a positive role in the development of UTCs they must be functional in the sense that they facilitate the development of scientifically rigorous theories of the information processing mechanisms that support human (and possibly animal) cognition. In other words, we seek criteria that may be applied descriptively to historical cases and prescriptively to facilitate scientifically progressive architectural development.

3 Lakatos' proposal: The cumulative research program

Popper was prescriptive in his approach. In his arguments for falsification rather than confirmation (Popper, 1935) and his later arguments for the use of bold conjectures to test theories (Popper, 1965), he was concerned with providing science with sound methods. Lakatos (1970) took a more descriptive approach. On the basis of historical analyses, Lakatos' argued that while the empirical nature of science was an important characteristic, real scientists did not discard their theories when their predictions were not met; rather, predictive failures were the catalyst for theory development. Furthermore, Lakatos argued, empirical testing of theories was not unguided and scientific theories did not develop in a haphazard way.

Lakatos (1970) argued that a theory in isolation could not be considered either scientific or pseudo-scientific. He took the view that science was characterized by its methods rather than by its theories, and hence that it was theory development, and the way in which a theory responded to predictive failures, that distinguished science from pseudo-science. He argued that theories, whether they are scientific or pseudo-scientific, consist of two parts:

- A *hard core* of central assumptions to which the theoretician is committed; and
- A *protective belt* of peripheral hypotheses that are the subject of theoretical advance.

[3] This option is not logically inconsistent with a family resemblance view of scientific theories: one could argue that while neither "being falsifiable" nor "being developed within a structured research program" are necessary characteristics of a scientific theory, both of these characteristics are shared by many scientific theories. It is not the purpose of this paper to explore this argument.

On Lakatos' view the theoretician makes a methodological decision that core assumptions are irrefutable. The *process* of science is then concerned with the resolution of anomalies and apparent counter-examples by appropriate adjustments to the peripheral hypotheses and incorporation of those hypotheses into the hard core. It is for this reason that such hypotheses are the subject of theoretical advance. The concept of a succession of theories, each with a hard core and protective belt, and with theoretical advance based on adjustments to peripheral hypotheses and the incorporation of peripheral hypotheses into the hard core, defines a Lakatosian research program.

While the concept of a Lakatosian research program emphasizes theory change, it does not discriminate between positive and negative change, or between scientific research programs and pseudo-scientific research programs. Lakatos (1970) also posited criteria for assessing change. According to these criteria, a research program is *theoretically progressive* if "each new theory has some excess empirical content over its predecessor" (Lakatos, 1970, p. 118) and *empirically progressive* if "some of this excess empirical content is also corroborated" (Lakatos, 1970, p. 118). A research program is *scientific* if it is at least theoretically progressive. Otherwise it is *pseudo-scientific*. It is *progressive* if it is both theoretically and empirically progressive, and *degenerating* if not. These distinctions are summarized in Table 1. A degenerating period within a scientific research program does not mean that the research program is doomed, as future empirical results may corroborate theoretical predictions resulting in a progressive research program. They are, however, cause for concern.

Table 1: Varieties of Research Program (based on Lakatos, 1970)

		Theoretically Progressive (increasing empirical content)	
		Yes	*No*
Empirically Progressive (predictions corroborated)	*Yes*	Scientific; Progressive	Pseudo-scientific
	No	Scientific; Degenerating	Pseudo-scientific

Lakatos' (1970) concept of a research program was intended to augment Popperian falsification rather than to replace it. Lakatos argued that scientists make what he referred to as a "methodological commitment" to their hard core of central assumptions: In the absence of any better methodology, scientists assume that their core assumptions are true and do not subject them to empirical testing. Falsification remains a key element of the methodology in the way that peripheral hypotheses support theoretical advance. The approach is effectively to attempt to falsify peripheral hypotheses in the context of a hard core of accepted assumptions. Lakatos argued that this restricted form of falsification is necessary; there is simply no alternative given that a) all observation is theory dependent (and so any data we use to corroborate or falsify a theory itself assumes that we are committed to a theory of observation), and b) science is a cumulative endeavor (and so science progresses by assuming a body of knowledge and building on that).

4 Cognitive architectures as Lakatosian research programs

Three significant implications follow if one accepts Lakatosian methodology as an appropriate model for the development of mechanistic theories of the cognitive architecture. First, theorists should distinguish between core assumptions and peripheral hypotheses. Second, theoretical development should be accompanied by empirical research focused on falsifying or corroborating peripheral hypotheses. Third, over time peripheral hypotheses should be incorporated into the hard core. These implications may be considered both in historical analyses of architecture developments to date, and as normative criteria for facilitating progressive development.

4.1 Central assumptions and peripheral hypotheses in Soar and ACT-R

Despite Newell's (1990) call, existing architectures have not explicitly adopted a Lakatosian methodology. Nevertheless, it is possible within the development of two of the most influential cognitive architectures of the last quarter century – Soar and ACT-R – to distinguish between central assumptions and peripheral hypotheses.[4] Within Soar, for example, one can argue that one central assumption is that cognition is goal directed and proceeds through the selection of *operators* which transform a representation of the problem state, while a second is that cognitive processing proceeds via a cyclic process consisting of an elaboration phase, where knowledge is brought to bear to propose and argue for or against alternative operators, followed by a decision phase, in which one operator from those proposed is selected (see, e.g., Newell, 1990, pp. 170–174). Details of the procedure used to weigh up alternatives within the decision phase would, however, appear to represent peripheral hypotheses. At least this was the case in the first half of the 1990s, when those details were revised on several occasions (see Laird & Rosenbloom, 1996).

The task of distinguishing between central assumptions and peripheral hypotheses in Soar is necessarily subjective because while Newell (1990) advocated a Lakatosian approach, Newell himself never provided a definitive list of central assumptions or peripheral hypotheses. Indeed Newell (1992) argued, in apparent contradiction with his earlier appeal to Lakatos, that no such list could be provided. The situation with respect to ACT-R is different. In his early work, Anderson was clear that the original ACT theory (the precursor to ACT-R) was built upon a set of explicit "preconceived notions":

> The shape of the ACT theory has been strongly influenced by preconceived notions […] about the nature of cognitive functioning. […] The biases are not precisely defined [… and they …] do not completely specify the ACT theory. They basically provide the skeleton of the model which [… is fleshed out …] with assumptions that would enable ACT to fit the data. When ACT proves wrong on some score, it is the 'fleshing-out' assumptions that are first to be sacrificed.
>
> (Anderson, 1976, pp. 114–115).

Anderson's "preconceived notions" are none other than Lakatosian central assumptions, while the fleshing-out assumptions that enable ACT to fit the data are Lakatosian peripheral hypotheses. The characterization of the theory in terms of preconceived notions fleshed out with additional assumptions, with the additional assumptions bearing the brunt of testing, appears to have been arrived at by Anderson (1976) independently of Lakatos (1970). It may be taken as further evidence for the relevance of Lakatosian criteria to the development of theories of the cognitive architecture.

Anderson's view captures the key principle of Lakatosian methodology as applied to the development of theories of the cognitive architecture. It is not simply that an architecture is a scientific theory that changes with time, or that by adopting Lakatosian methodology architectures do not need to be accompanied by an empirical strand of research. The key principle is that by differentiating between central assumptions and peripheral hypotheses Lakatosian methodology gives guidance on the design of empirical tests for the theory, and guidance on how to respond to those tests. From a Lakatosian perspective, empirical tests should test peripheral hypotheses. If predictions are met, these empirical tests give support to peripheral hypotheses, which may eventually be incorporated into the theory's hard core of central assumptions. If predictions are not met, it is the peripheral hypotheses which must be adjusted to bring the theory in line with data.

[4] See Cooper (2006) for a reconstruction of Soar and ACT-R as Lakatosian research programs.

4.2 Theory development in Soar and ACT-R

While both Soar and ACT-R may appear Lakatosian in that central assumptions and peripheral hypotheses may be identified for each of them, the critical element for distinguishing science from pseudoscience for Lakatos was theory development. In this section we therefore consider some critical developments in Soar and ACT-R from a Lakatosian perspective. To anticipate, both architectures have witnessed theory change that from a Lakatosian perspective would be classified as scientific and progressive, but both have also been subject to more problematic theory change.

4.2.1 The case of Soar

Soar developed from Newell's early interest in human problem solving (see, for example, Newell & Simon, 1972). The basic architecture consists of a working memory which contains a representation of details of the current goal and task and a long-term memory containing condition-action rules. Processing is cyclic with each cycle consisting of an elaboration phase, in which all rules are applied to the contents of working memory to generate a set of possible operators (i.e., ways of transforming the problem solving state) and preferences indicating the relative or absolute worth of those operators, followed by a decision phase, where the most preferred operator is selected and installed in working memory. The next cycle is then initiated. If processing is blocked because, for example, no operators are available or multiple operators appear equally suitable, Soar will automatically create a subgoal to explore different possible courses of action. Processing then shifts to this subgoal. When the subgoal is resolved (through Soar deciding upon a single preferred operator at the higher level), Soar automatically creates a new condition-action rule whose conditions are the relevant aspects of the processing state that lead to the blockage and whose action is to prefer the operator that resolved the blockage. The blockage is known as an impasse, the creation of the subgoal is known as automatic subgoaling, the new condition-action rule is known as a chunk, and the process of creating the rule is known as chunking (see Newell, 1990, for more details).

The basic mechanism of universal subgoaling followed from a fundamental desire of the Soar theorists to have the architecture set its own subgoals. This led to the proposition that all subgoals arise from impasses in problem solving and the subsequent classification of abstract impasse types (including tie impasses, where multiple operators appear to be equally good, and no-change impasses, where no operators appear suitable). The basic classification (Laird *et al.*, 1987) has withstood the test of time. The chunking learning mechanism, which has also withstood the test of time, follows directly from universal subgoaling, in that once the architecture can create and resolve subgoals in response to impasses, it is natural to record the conditions that led to the impasse and its resolution within a new condition-action rule. These advances in the Soar theory were broadly successful, allowing Soar to be applied with some success to a range of tasks. The advances are theoretically progressive in the Lakatosian sense (as they resulted in an architecture with greater scope and hence greater empirical content), but assessment of the developments as empirically progressive depends on the stance one takes on prediction. It is unclear, for example, if power-law speed up in behavior following learning in Soar was an *a priori* prediction of the chunking mechanism or a *post hoc* finding which matched the existing literature. Either way, these early developments in the history of Soar would seem to be scientific in Lakatos' terms. This is despite that fact that the details of the universal subgoaling and chunking mechanisms changed in almost every release of Soar since 1987. In Lakatosian terms such details are peripheral hypotheses, and changes to such hypotheses do not impact upon the scientific status of the theory (provided that such changes do not reduce the empirical content of the theory).

A more problematic development in Soar's history was the introduction of the single state principle, in the guise of "destructive state modification". Prior to this development, Soar could support multiple states within a single representation of a problem. Thus, a problem

representation could simultaneously include one state in which an object was closed and another in which the same object was open. This was considered to be implausible for well-grounded psychological reasons (see Newell, 1990, pp. 251–252). Furthermore, limiting the architecture to a single state for each problem representation yielded an explanation for one of the most common human problem solving strategies, that of progressive deepening (Newell & Simon, 1972). The single state principle was therefore incorporated in Soar 5 in 1990. That incorporation took the form of destructive state modification – a set of mechanisms that allowed the state of a problem representation to be modified. While this is justifiable in theory, the implementation introduced numerous complications (such as micro-decisions on almost all elements of working memory and mico-impasses relating to those decisions). Many of these complications have since been reversed.

The apparent failure of destructive state modification (evidenced by successive retraction of changes introduced in Soar 5) may be attributed to at least two factors. First, the developers of Soar made an early commitment to a specific (and overly complex) instantiation of the principle in the apparent absence of any attempt to explore alternative instantiations. Alternative instantiations involving fewer alterations to Soar 4 clearly existed. The current implementation of Soar (Soar 8) contains one, while another was suggested by Cooper *et al.* (1996).[5] Second, the implementation of destructive state modification in Soar 5 interacted in complex ways with the assumptions of earlier versions of the theory (particularly in relation to learning, but also in relation to more basic aspects of processing such as selection and application of state-changing operators). This made it difficult to attribute predictive failures to specific peripheral hypotheses.

4.2.2 The case of ACT-R

ACT-R's heritage lies in Anderson and Bower's early work on human associative memory (Anderson & Bower, 1973), but the architecture shares some features with Soar. First, like Soar there is a distinction between two memory stores: a procedural memory (whose elements are condition-action rules) and a declarative memory (which serves a similar function to Soar's working memory). Processing, as in Soar, involves the application of condition-action rules, but selection of which rule to fire at any time is based on a numerical calculation that weighs the costs and the anticipated benefits of each applicable rule: on each processing step, all rules that match the current goal are considered and that with the greatest utility is selected. Unlike most versions of Soar, elements in declarative memory have associated activation values, with activation spreading between associated elements. The time taken to apply a rule, once selected, depends on the time taken to match or retrieve its conditions in declarative memory, which in turn depends on the level of activation of declarative memory elements.

Since the early 1990s, work within the ACT framework has concentrated on revising the architectural mechanisms so that they are consistent with the principle of rationality, namely that "the cognitive system optimizes the adaptation of the behavior of the organism" (Anderson, 1991, p. 3). In Lakatosian terms, this move has clearly been both theoretically and empirically progressive, but this move did not occur overnight or in one step. Thus, Anderson's original arguments for rational analysis (Anderson, 1990) were not tied specifically to the ACT framework. Indeed, Anderson (1990, p. xi) made it clear that there appeared to be some tension between rational analysis and the architectural approach. A partial resolution of this tension came three years later in the form of principled equations for determining how the activation of declarative memory elements and the numerical parameters underlying the utilities of condition-action rules change through experience (Anderson, 1993). Adjustment of the rational analysis aspects of ACT-R continued throughout the 1990s as refinements to the equations were developed and explored (see Anderson & Lebiere, 1998, pp 431–439). The incorporation of

[5] It is possible that such alternatives were not explored in the early 1990s because of prevailing sociological factors, namely Newell's declining health.

rational analysis into the ACT framework thus illustrates that an exploratory/experimental approach to theory development can bear fruit. It is, however, both time and resource intensive.

A more problematic example of theory evolution within the ACT series concerns the account of rule learning. Successive versions of ACT* and ACT-R have incorporated a range of different approaches to the learning of new condition-action rules. Thus, ACT* initially supported four mechanisms by which condition-action rules could be learned – *discrimination*, in which the conditions of an existing rule are augmented to limit the rule's applicability, *generalization*, in which the conditions of an existing rule are relaxed to allow the rule to apply more generally, *composition*, in which two rules that are performed in sequence could be combined into a single rule, and *proceduralization*, in which the actions of a rule could be specialized, thereby eliminating the need to recall declarative knowledge. Discrimination and generalization were removed from the 1987 version of the ACT* theory due to problems in limiting their applicability and lack of empirical support (Anderson, 1987). The initial version of ACT-R (Anderson, 1993) replaced the remaining mechanisms with a single mechanism based on analogy (allowing the formation of new condition-action rules by analogy with existing rules). This was then replaced with a production compilation mechanism (Anderson & Lebiere, 1998) that bore some similarities to Soar's chunking mechanism in that learning was linked to the creation and resolution of subgoals. This too has now been replaced. The current approach (as described by Taatgen & Anderson, 2002) allows memory retrieval operations to be merged into existing condition-action rules that subsequently use the retrieved information. It shows considerable promise, but it is too early to say whether it has resolved all outstanding issues.

Why has learning within the ACT framework been subject to so much flux? Unlike other aspects of ACT-R, there is a tight integration or dependence between possible learning mechanisms and the rest of the architecture. Thus in the current instantiation (ACT-R 6.0) learning is intimately related to the use of ACT-R's goal buffer, a construct extensively revised in the transition from ACT-R 4.0 to ACT-R 5.0. Clearly, some level of dependency is inevitable for a mechanism such as learning, which can always be thought of as an add-on that modulates the functioning of the system by creating or modifying long-term knowledge structures, rather than a more "basic" mechanism of central importance such as, say, production matching. However, dependencies between assumptions are relatively rare within ACT-R in that one can explore variants of ACT-R in which different assumptions are swapped in or out. For example, within the basic activation-based production-system substrate, one can explore different approaches to the flow of activation, the matching and selection of productions, the creation and maintenance of goals (cf. Altmann & Trafton, 2002) and even the creation of new productions and the tuning of production parameters during processing (i.e., learning).

The dependency of learning assumptions on prior assumptions is therefore relatively unusual in ACT-R. It has two consequences. First, assumptions relating to learning are at the mercy of more basic (in the sense of the previous paragraph) assumptions, and second, the consequences of learning cannot easily be attributed to the specific learning assumptions, for they result from the interaction of the learning assumptions and the more basic assumptions on which learning is founded. This second consequence echoes the difficulties faced by Soar following introduction of destructive state modification. However both of these consequences work against the smooth development of ACT-R's learning mechanism.

4.3 Contrasting approaches to theory development

The Lakatosian analysis of developments in Soar and ACT-R ignores the fact that radically different approaches to theory development have been adopted by theorists working with Soar and ACT-R. Thus, when confronted with an apparent predictive failure, the approach of Newell and the Soar community, at least during the 1980s and 1990s, was one of "listening to the architecture" (Rosenbloom, et al., 1993, p. xxv). This approach was complemented by a methodological assumption of one mechanism for each function (and so, for example, a single mechanism for all forms of learning). The Soar community would therefore attempt to develop

solutions to predictive failures by using the existing architecture in novel or ingenious ways. For example, learning of paired associates presented difficulties for Soar. The basic task requires that, when given a nonsense cue (e.g., BAJ), the architecture should produce the associate given in an earlier training session (e.g., GID). However, the obvious way of applying Soar's learning mechanism to the problem led to a situation in which Soar could only produce an associate if given both the cue and associate, not just the cue (Newell, 1990, pp. 326–345). Following detailed analysis of learning within the architecture, the paired-associates problem in Soar was solved by breaking the task in the training session into generate and test subtasks, with the generate subtask producing all possible associates and the test subtask picking out the correct one. Thus, rather than modify the architecture, the architecture was analyzed to understand the reason for the difficulty with learning paired associates and how that difficulty could be overcome.

The contrast with ACT-R is clear. Experimentation with the mechanisms and assumptions underlying ACT-R is common and encouraged. For example, Byrne & Anderson (2001) were concerned with perceptual and motor aspects of behavior, an area that had previously been outside of the scope of the ACT theories. Their inspiration was the work of Meyer & Kieras (1997a, 1997b), who developed EPIC, a novel architecture in which perceptual and motor systems were assumed to be distinct, modular, subsystems that interacted with a central parallel processor. Building on this, Byrne & Anderson (2001) developed ACT-R/PM, a variant of ACT-R in which a central serial processor (basically the standard ACT-R system) interacted with numerous modular perceptual and motor subsystems. ACT-R/PM was shown to be capable of capturing much of the data which had been cited in support of EPIC. In addition, it made novel predictions about so-called "cognitive PRP effects" – predictions which Byrne & Anderson (2001) went on to test. After additional experimentation with other tasks, the modularization of ACT-R/PM was incorporated into the main ACT-R theory.

The ACT-R community thus supports an approach that has been summarized as "let a thousand flowers bloom" (Anderson, 2003, personal communication). If one measures the success of an architecture in terms of the number of researchers working with it or the number of published peer-reviewed papers in which it serves an explanatory role, then ACT-R is currently more successful than Soar. The relative success of ACT-R may be attributed in part to this "let a thousand flowers bloom" approach and the relative independence of ACT-R's assumptions. The first of these emphasizes the empirical nature of ACT-R – exploring numerous variant architectures before committing to any one option – while the second simplifies the attribution problem.

In fact, in recent years a more experimental approach appears to have been adopted by some in Soar community, with renewed interest in Soar being accompanied by a range of proposed augmentations to the architecture, including an EPIC-like perceptual-motor component (Chong & Laird, 1997), activation-based working memory (Chong, 2003; Nuxoll et al., 2004) and a reinforcement learning mechanism (Nason & Laird, 2004). It is far too soon to attempt an evaluation of the long-term impact of these developments, but the shift towards a policy of experimenting with the architecture reflects a fundamental change in the approach to architecture development by proponents of Soar.

5 Implications

We have seen that Lakatos' characterization of theories in terms of a hard core of central assumptions and a protective belt of peripheral hypotheses, together with his proposed characterization of the process of science in terms of a cumulative methodology whereby peripheral assumptions are absorbed into the hard core, provide a plausible model for the development of theories of cognitive architecture. The model does not address the lack of falsifiability of cognitive architectures, however, as that lack of falsifiability arises from the necessity of task knowledge when developing predictions, and not from the existence of

peripheral hypotheses. It is legitimate to ask, then, if the Lakatosian perspective can serve any functional purpose in the development of cognitive architectures, or whether it simply offers the endeavor a smokescreen of scientific credibility. In this section it is argued that adoption of the Lakatosian perspective as a normative methodology for the development of a cognitive architecture can provide guidance not just on the development of the cognitive architecture, but also on the development of models that are embedded within the architecture and on how the cognitive architecture itself might be evaluated.

5.1 Prediction, falsifiability and attribution

Lakatosian methodology distinguishes between falsifiability and prediction. Predictive failures do not falsify central assumptions. They falsify peripheral hypotheses in the context of central assumptions. Thus, in order to address predictive failures within a Lakatosian research program, it is necessary to attribute each predictive failure to one or more specific peripheral hypotheses. The relevant hypotheses may then be adjusted in an attempt to address the predictive failure. There are two key difficulties in this process of addressing predictive failures: attribution and hypothesis adjustment. The latter is an issue for the specific architecture. The former is more generic.

Before considering the problem of attribution, it should be noted that predictive failures are not necessarily bad, as they provide the impetus for theory development. Thus, if a cognitive architecture is found never to fail (possibly because testing is too conservative or because the architecture is insufficiently constraining), then there will be no pressure to develop the architecture and it will stagnate. At the same time if theory development were unduly responsive to predictive failures then development would be ad hoc. Thus, predictive failures are positive but must be understood (and prioritized) within the context of long-term research goals (i.e., within the context of the "positive heuristic" as discussed below).

Given that a predictive failure has occurred, it is far from obvious, a priori, that attribution will be unambiguous. In general, cognitive architectures consist of multiple central assumptions and multiple peripheral hypotheses that interact in yielding a behavior. Thus, when a cognitive architecture's behavior differs from observed behavior, attribution requires isolating one or more specific peripheral hypotheses whose interaction with central assumptions results in the predictive failure. Attribution will be most straightforward when peripheral hypotheses are tested in isolation from each other. Experimental testing of an architecture therefore needs to be conducted with attribution in mind.

The situation is complicated by the fact that, as noted above, when a cognitive architecture is supplemented with task knowledge in order to develop predictions, predictive failures may be attributed either to architectural peripheral hypotheses or to task peripheral hypotheses. It is tempting to suggest that, where attribution is ambiguous, that preference should be given to task peripheral hypotheses, and these should be the subject of revision rather than architectural peripheral hypotheses. This line of argument is justified by the fact that modification of architectural peripheral hypotheses is likely to have implications for other models developed within the architecture, and so these should only be modified as a last resort. However, a more progressive strategy is to consider architectural development and task development as separate research programs. This has significant implications for the use of cognitive architectures in developing task models (as discussed below in the section *Model Development within an Architecture*), but this second strategy appears to have emerged naturally in the case of the ACT-R architecture.

The strategy of separate research programs for cognitive architectures and task models is only feasible if peripheral hypotheses from both cognitive architectures and task models can be tested in isolation from each other. Examination of one recent modification to the ACT-R architecture, the replacement of the goal stack by a goal buffer with activation-based retrieval of goals from declarative memory, suggests that this can be the case.

Goal-directed processing was introduced within the ACT series of architectures in Anderson's (1983) presentation of ACT*. It was not present in the original ACT (Anderson, 1976) or HAM (Anderson & Bower, 1973) theories from which ACT* evolved. It is clear that Anderson was never comfortable with the concept of an unlimited goal stack as adopted by the Soar architecture (Laird, et al., 1987; Newell, 1990). Thus, in the original description of ACT*, Anderson stated that "it might be reasonable to assume that three to five goals are available … The rest of the goals, however, will have to be retrieved from long-term memory if they are to be had at all" (Anderson, 1983, p. 161; see also Anderson, 1993, p. 136, for comments concerning the goal stack in ACT-R 2.0, and Anderson & Lebiere, 1998, p. 459, for comments concerning the goal stack in ACT-R 4.0). Nevertheless, an unlimited goal stack was implemented as a convenience within the computational instantiations of the ACT* theory and all versions of its successor, ACT-R, prior to 2003.

In considering the status of goals within the problem solving literature, Altmann and Trafton note that "a reader might be forgiven for coming away with the view that backtracking was supported by some kind of special cognitive structure" (Altmann & Trafton, 2002, p. 43; see also Altmann & Trafton, 1999). A goal stack is of course the specialized cognitive structure to which Altmann and Trafton refer. Altmann and Trafton (2002) go on to criticize not just the concept of a goal stack, but the concept of a dedicated goal memory, and develop a model of the Tower of Hanoi task (a goal-subgoal intensive task) within the ACT-R architecture using only general spreading-activation processes for goal retrieval. The model is shown to perform as well as an existing ACT-R model (that of Anderson & Lebiere, 1998) on some measures (move latencies), but it does this with fewer assumptions. In addition, the Altmann and Trafton model is able to account for error data, on which the Anderson and Lebiere model is silent. Altmann and Trafton (2002) therefore provide specific, targeted, evidence against the necessity of the goal stack assumption, while at the same time providing a plausible alternative consistent with other assumptions of the ACT-R architecture. Furthermore Altmann (2002) and Altmann and Gray (2002) provide specific experimental evidence for the decay of goal representations (and the need for deliberate "refreshing" of goal representations) in tasks beyond the Tower of Hanoi. This evidence is itself consistent with the approach of Altmann and Trafton (2002) and inconsistent with a special-purpose goal stack. Finally, Anderson & Douglass (2001) present experimental evidence from a variant of the Tower of Hanoi task designed specifically to address the issue of goal encoding and retrieval costs. Their data suggest that goals are subject to the same decay processes as other information in declarative memory, and moreover that, at least in their variant of the Tower of Hanoi task, participants are able to reconstruct goals if they are not accessible (e.g., because their trace has decayed). These findings hence also argue against a special purpose goal stack.

This case study demonstrates that within ACT-R it is possible to target individual peripheral hypotheses. It is not clear that this targeting is possible in all cognitive architectures. As discussed above in the section *Theory Development in Soar and ACT-R*, ACT-R is a somewhat loose collection of independent assumptions and hypotheses. Within Soar, for example, hypotheses are more tightly integrated and targeting one assumption independently of others may require greater ingenuity.

5.2 From data fitting to prediction

The preceding section assumed that predictive failures may occur within an architecture's research program. This requires some comment as the view that behavior only follows from an architecture when parameters are fixed, a task theory is adopted, and implementation assumptions (if present) are made concrete suggests that true prediction is not possible.

In general, task theory assumptions, parameters and architecture implementation details each contribute degrees of freedom to the behavior to be modeled, and true prediction only follows when the number of such degrees of freedom is zero. In fact, qualitative predictions can be made even with non-zero degrees of freedom. For example, Soar's chunking mechanism

predicts that learning in all tasks will follow a power law. At the same time, both Soar and ACT-R research programs have made clear attempts to reduce the number of degrees of freedom as they have developed. Thus, Soar has few parameters but attempts have been made to fix the timings of various sub-processes (see Newell, 1990), while in ACT-R, which has of the order of 30 parameters (see Table 12.3 of Anderson & Lebiere, 1998, pp. 434–435), default values have been established for all parameters and these default values are routinely used across tasks. There is also an attempt to develop models within ACT-R that process task instructions, thereby reducing the degrees of freedom available in coding specific task theories within the architecture (Anderson, 2006, personal communication; see also Huffman et al., 1993, for a related attempt within Soar).

All of these tactics reduce the number of degrees of freedom and therefore allow for something approaching true, quantitative, prediction. Note though that even without these methods of reducing degrees of freedom, it is still possible to achieve prediction through cross-validation. This technique, which is standard in machine learning, involves dividing the data to be explained into two subsets, and then fitting the behavior of the model + architecture to one subset though adjusting parameters. The resultant model + architecture can then be used to predict, with fixed parameters, the second subset of the data. The process may be repeated with different randomly selected subsets of the dataset. Cross-validation may be particularly important early in an architecture's development, when the number of degrees of freedom is likely to be high.

5.3 Model development within an architecture

Cognitive architectures have two distinct roles: as theories of the structure and functioning of the subprocesses supporting cognitive function, and (in virtue of this and the now standard method of specifying an architecture in terms of an implementation) as computational systems within which theories of behavior on specific tasks (ranging from the Stroop task to air traffic control tasks) or of specific cognitive functions (such as short-term memory, attention, or problem solving) may be developed. A Lakatosian approach to architectural development has implications for the development of models within that architecture. Most clearly, if a model's behavior depends on an architecture's peripheral hypotheses and the model is successful (i.e., it accounts for the relevant data) then this adds weight to the architecture's peripheral hypotheses. Equally, however, if the peripheral hypotheses are later revised or abandoned this will undermine the original model. It is therefore critical when developing a model within a cognitive architecture to understand the extent to which the model's behavior is dependent on the architecture's central assumptions and peripheral hypotheses.[6]

One illustrative case where dependence and independence can be seen arises in the work of Goal *et al.* (2001), who developed a model of the Tower of Hanoi task within the 3-CAPS architecture. They then used their model to investigate theories of frontal lobe dysfunction. While the model was developed within a specific architecture, the authors expressed minimal commitment to that architecture. For the purposes of the model, all that was important was that 3-CAPS provided a capacity limited working memory which could be impaired (to mimic frontal lobe dysfunction) by increasing the rate at which working memory elements decayed. Two questions follow: Would the results transfer to 4-CAPS (Just et al., 1999), the successor of 3-CAPS, and would the results transfer to other architectures with similar working memory subsystems? Both questions may be answered in the affirmative. The critical feature of 3-CAPS is retained in 4-CAPS, while Goel *et al.* (2001) specifically address the issue of development of the model in ACT-R, arguing that their results are consistent with the work of Kimberg and Farah (1993), who developed a general model of frontal lobe dysfunction in an early version of ACT-R. In sum, the work demonstrates that architectures can provide a useful substrate for the

[6] If a model's behavior is independent of both the architecture's central assumptions and its peripheral hypotheses then there is nothing gained by developing the model within the architecture – the architecture is serving as nothing more than an implementation language.

implementation and development of specific models, but that analysis of the dependency between model behavior and specific architectural assumptions is critical in order to draw out the full implications of the model.

5.4 Heuristics for theory development

Two additional concepts introduced by Lakatos (1970) but not discussed above are those of positive and negative heuristics for theory development. These are rules of thumb about what is, and what is not, permissible within Lakatos' model of scientific enquiry. The negative heuristic is constant across all research programs. It simply states that central assumptions are considered to be irrefutable, and hence that empirical work should not attempt to falsify these assumptions. The positive heuristic, in contrast, specifies a plan for theory development. It is "a partially articulated set of suggestions or hints on how to [...] modify, sophisticate, the 'refutable' protective belt" (Lakatos, 1970, p. 135).

The positive heuristic consists of a recipe for theory development. Lakatos (1970, pp. 135–136) cites Newton's approach to his theory of planetary motion as a prime example, arguing that Newton began by working out his theory with a point-like planet orbiting a fixed point-like sun. This was then modified such that the bodies orbited a common center. Multiple planets were then considered, and then the point-like bodies where generalized to massive bodies, and so on. Each stage required greater mathematical sophistication, and Newton even had to develop new mathematical techniques (the calculus) to support the theoretical development. The moral of this example is that successful scientific theories do not develop in an ad hoc fashion or even in a way that is excessively responsive to predictive failures. Thus, from a Lakatosian perspective if a cognitive architecture is to be methodologically sound it must be developed within an overall plan that is at least partially specified in advance.

The simplicity of the negative heuristic is misleading, for it has implications for when a theory (or at least some of its central assumptions) should be abandoned. This is when a theory's predictive failures can no longer be addressed through adjustments to peripheral hypotheses. While it is not in general possible to prove that any specific predictive failure cannot be addressed through some ingenious adjustment to one or more peripheral hypotheses, the weight of predictive failures and the increasing complexity of peripheral hypotheses needed to patch up the theory can eventually undermine a theory. A good case in point is learning within the Soar architecture. A central assumption of Soar prior to 2004 was that all learning arises from a single mechanism: the chunking of subgoal hierarchies which is held to occur continuously during all cognitive activities.

In its basic form, chunking is an elegant and powerful learning mechanism. As noted above it accounts for the so-called power-law of learning, and its hypothesized ubiquitous nature is consistent with the ubiquitous nature of the power-law of learning, which has been demonstrated in tasks ranging from choice reaction time (Seibel, 1963) to reading of inverted text (Kolers & Perkins, 1975) and proving theorems in a syntactic calculus (Neves & Anderson, 1981). Chunking does, however, have serious difficulties as the *only* form of learning. Thus, as discussed above rote learning of paired associates has long presented a major difficulty for the mechanism, as has learning from external feedback. Methods have been proposed which circumvent these difficulties, but the complexity of the methods (which effectively consist of peripheral hypotheses about the cognitive processes involved in learning in these situations) undermine the elegance of Soar's basic chunking mechanism. The predictive failures, together with the complex peripheral assumptions proposed in response to those failures, suggest that either Soar as a theory of the human cognitive architecture should be abandoned, or at the very least that Soar's theory of chunking as the only mechanism of learning must be reconsidered.

In fact, proponents of Soar have recently begun to explore the implications of the second of these possibilities. Thus, Nason & Laird (2004) have suggested augmenting the Soar architecture with a second learning mechanism (reinforcement learning: Sutton & Barto, 1998).

The rationale for this was not the difficulties in chunking highlighted above, but rather difficulties in learning within knowledge-lean statistical environments where feedback is limited to knowledge of success or failure. In addition, the incorporation of reinforcement learning is consistent with a wider development that has served to revitalize interest in Soar as a theory of cognitive processing, namely the incorporation of sub-symbolic, activation-based, processing within working memory (Chong, 2003; Nuxoll et al., 2004).

Together, the positive and negative heuristics allow us to answer a central question in the development of a cognitive architecture: How can one maximize progressive development while guarding against false advances? The Lakatosian answer lies in articulating central assumptions, acknowledging predictive failures, and having a clear plan for theory development in which those predictive failures may be addressed. In the case of learning within Soar, the central assumptions were well articulated, and the predictive failures were clear. The difficulty for Soar prior to 2004 was that there was no plan in which predictive failures might be addressed beyond sophisticating the peripheral hypotheses in whatever way was necessary to preserve the assumptions of the hard core. Thus, there were few if any constraints on how the peripheral hypotheses might be modified or extended to account for the problematic learning phenomena.

Could things have been done differently? One can envisage alternative approaches to the problems of chunking within Soar, such as introducing additional learning mechanisms or even modifying the conditions under which subgoals are created or resolved, but these all involve rejecting central assumptions of the Soar architecture, and without an overall plan for theory development would in any case be ad hoc. Thus, the main difficulty for Soar in responding to the predictive failures of chunking was that Soar's positive heuristic which had operated since the early 1980s had been exhausted. While that heuristic had not been well articulated, it guided research within Soar for well over a decade via the addition of a range of concepts and mechanisms, including impasses and automatic subgoaling, chunking, and the single state principle.

To suggest that the positive heuristic which drove Soar development through the 1980s and 1990s has been exhausted does not mean that Soar was or is a failure. At the time of writing interest in Soar appears to be growing, possibly because a new positive heuristic now appears to be driving theory development (as reflected in the incorporation of activation-based processing in Soar's working memory). Regardless of this renewed interest, Soar has a significant legacy even if it is rejected as a theory of the human cognitive architecture. Below we suggest (following Lakatos, 1970) that unified theories of cognition be judged on their "heuristic power". Soar's positive contribution in this respect is great.

5.5 Heuristic power and the evaluation of architectures

A final critical issue concerns the evaluation of architectures. The Lakatosian concept of heuristic power provides a way in which the contribution of an architecture's research program may be quantified. Lakatos defines heuristic power rhetorically by asking of research programs (or the positive heuristic that has driven them) "how many new facts did they produce, how great was their capacity to explain their refutations in the course of their growth?" (Lakatos, 1970, p. 137). What counts as a "new fact" is not always clear,[7] and the difficulties inherent in falsifying architectures undermines the force of the second of these questions, but we may still ask the first of these questions of Soar and ACT-R. Indeed, we may ask the same question of the connectionist research program revived by Rumelhart and McClelland (1986; see also McClelland & Rumelhart, 1986).

Consider first the case of Soar. Evaluation of Soar is compromised by difficulty in identifying its positive heuristic. As suggested above, since 1990 Soar development has largely been driven

[7] For example, does a novel account of an existing fact, not previously considered in the development of the architecture, count as a "new fact"?

by incorporation of the single state principle and real-time AI concerns. Only recently has a group of researchers working with Soar returned to its psychological roots. While the single state principle had a solid psychological foundation, it has not produced any obvious new facts. Development related to real-time AI concerns has been based purely on engineering concerns, and so it too cannot provide positive evidence of Soar's heuristic power. It is too early to judge the success of the current wave of psychological development of Soar.

The evaluation of ACT-R, and particularly rational analysis, is more positive. In part, this is because the ACT-R research program has retained a strong empirical strand, but this is not the only reason. Anderson (1990) demonstrated that rational analysis could account for aspects of behavior across four domains (memory, categorization, causal inference and problem solving), and the incorporation of rational analysis into the ACT framework has led to the colonization by ACT-R of numerous additional domains within cognitive psychology (e.g., choice, cognitive arithmetic and analogy: see Anderson & Lebiere, 1998). In short, the heuristic power of rational analysis, particularly in combination with the ACT framework, has been great.[8] Evaluation of ACT-R's most recent development, driven by modularization and mapping to neurophysiology, also appears to be positive. Again, these have produced genuine predictions which have, following subsequent experimentation, been corroborated (e.g., Byrne & Anderson, 2001; Anderson et al., 2004; Anderson, et al., 2005).

Turning to connectionism, it is hard to identify a single positive heuristic that has guided the approach. Nevertheless there are numerous examples were connectionism has provided novel accounts of phenomena, particularly within neuropsychological modeling (e.g., Hinton & Shallice, 1989). On these grounds, connectionism has been a successful research program. Connectionism is a doubly interesting case as it has been subject to numerous attempted refutations (e.g., Broadbent, 1985, Fodor & Pylyshyn, 1988, and subsequent articles). The extent to which connectionist attempts to explain refutations (e.g., Chater & Oaksford, 1990; Elman, 1990) were successful remains a matter of debate, but it is clear that the positive heuristic guiding the development of connectionist models is also changing, with increased attention now being directed at functionally modular models with increased neurophysiological plausibility (e.g., O'Reilly & Munakata, 2000).

Given all of the above, an architecture is successful to the extent that it remains theoretically or empirically progressive. From this perspective, while Soar may have been successful during the late 1980s and early 1990s, architectural developments following the publication of Newell's call to arms (Newell, 1990) were neither theoretically nor empirically progressive. More recent developments may, in time, yield a more positive assessment. In contrast, ACT-R is currently highly successful precisely because it remains theoretically and empirically progressive. Other less high profile architectures may also be judged on the criteria of theoretical and empirical progressiveness, though that judgment is hampered by the scarcity of publications relating to other architectures.

Is ACT-R likely to remain successful? The answer to this largely depends on whether there is much mileage left in the positive heuristic that has driven ACT-R through much of the last fifteen years (rational analysis), and whether a new positive heuristic can be found to further theory development once rational analysis is exhausted. Two related recent developments in ACT-R mentioned above, namely the modularization of non-central processes and the subsequent mapping of modules to neural structures, suggest that a new positive heuristic, based on a concern for neuroscience, has emerged. The immediate future of ACT-R therefore appears assured. This does not mean that other, "competing", architectures will continue to be

[8] The substantial heuristic power of rational analysis raises the intriguing question of whether rational analysis could have been adopted within the Soar research program, and if so whether it might have been as successful.

dominated by ACT-R, particularly if such architectures successfully develop and exploit their own positive heuristics.

6 Discussion and conclusion

The issue of architecture development has been couched in terms of Lakatosian research programs because of Newell's explicit reference to Lakatos and Anderson's explicit but independent adoption of a highly similar methodological position. For these reasons, Lakatos' account of scientific progress seems particularly apt in the case of cognitive architectures. At the same time Lakatos' account is not the only account of theory change within science. We briefly consider the relevance of two more recent accounts.

Laudan (1977) argued that in developing his account of research programs Lakatos (1970) worked with an idealized view of history. According to Laudan, real scientific theories – even the ones cited by Lakatos – do not develop according to the strict rules of Lakatosian research programs. The above consideration of the development of Soar and ACT-R is consistent with this argument – there is evidence in both cases of non-progressive development. Laudan argued that Lakatos (1970), and Kuhn (1962) before him, both falsely viewed theoretical development in strictly monotonic terms (e.g., in the case of Lakatos, through an increasing number of central assumptions between one instance of a theory and the next). Laudan also argued that Lakatos and Kuhn both failed to consider the role of conceptual problems in the development of theories. Both Kuhn and Lakatos accepted logical positivist notions of theoretical progress, in which progress was defined in terms of one theory having greater empirical content than its predecessor. Thus, neither account allowed for progress that might come from reducing a theory's conceptual problems (e.g., by removing an ad hoc assumption), unless that reduction also led to greater empirical content.

Laudan's difficulty with monotonicity is well illustrated with reference to theories of the cognitive architecture through the recent exploratory work involving the abandonment of some of Soar's hitherto central assumptions (chunking as the only learning mechanism and the all-or-nothing representation of elements in working memory), and this move within Soar might be taken as support for a view of architecture development based on the concept of a *research tradition*, which Laudan (1977) presented as an alternative to Lakatosian research programs. This position has several immediate difficulties. First, research traditions are characterized by "general assumptions about the entities and process in a domain of study, and about the appropriate methods to be used for investigating the problems and constructing the theories in that domain" (Laudan, 1977, p. 81). Thus, behaviorism, information processing psychology, and even computationalism might be viewed as research traditions, but specific architectures such as Soar or ACT-R are not. Instead, instances of those architectures (e.g., Soar 4.5 or Soar 8.2 or ACT-R 4.0 or ACT-R 6.0) are more accurately viewed as theories within a research tradition grounded in computationalism. Second, in abandoning monotonicity Laudan accepts that scientific progress need not be cumulative, but in so doing he sacrifices the distinction between central assumptions and peripheral hypotheses. Together, these two difficulties imply that to apply Laudan's approach to the development of a cognitive architecture would require greater specification of the relation between successive theories within a research tradition, and of how competing theories within a research tradition may be closely or distantly related.

A final concern raised by Laudan's position is that he provides no account of how science might be demarcated from pseudo-science. In later work Laudan goes even further, denying that demarcation is an issue and arguing that "we ought to drop terms like 'pseudo-science' and 'unscientific' from our vocabulary" (Laudan, 1983, p. 125).

In the light of these points, whole-hearted acceptance of Laudan's approach cannot proceed without elaboration of the relation between theories within a research tradition, and even given that, Laudan's rejection of demarcation is unlikely to address the concerns of those who criticize

cognitive architectures on the grounds that they are not falsifiable. At the same time, Laudan's notion of progress – through maximizing the number of empirical problems solved while minimizing the number of conceptual problems raised – cannot be easily dismissed.

A second alternative to the Lakatosian view is provided by Thagard's (1992) work on conceptual change and explanatory coherence. Thagard's distinctive contribution was to develop and apply cognitive scientific theories of knowledge and conceptual representation to the analysis of scientific change. Thagard uses a constraint satisfaction algorithm to model the relations between evidence and hypotheses – and hence the "explanatory coherence" – of a number of influential scientific theories (including many that have been superseded by theories with greater explanatory coherence). In principle the approach might be applied in a descriptive fashion to chart the development of Soar or ACT-R. Thagard's approach is of interest, particularly in the light of the difficulties raised with Laudan's concept of research traditions, because his simulations allow one to quantify the degree to which individual hypotheses contribute to the explanation of a body of evidence. It thus offers the prospect of allowing one to reconstruct a graded version of the central/peripheral distinction of Lakatos, without committing to a strictly cumulative view of theory development.

Notwithstanding these alternate characterizations of scientific progress, we have seen that the concept of a Lakatosian research program offers a plausible model for the development of cognitive architectures through the distinction between central assumptions and peripheral hypotheses, but that this model does not directly address the issues raised by the lack of falsifiability of cognitive architectures. We have argued, however, that even without a response to the falsifiability issue, adoption of the Lakatosian model in the development of cognitive architectures has numerous advantages. Thus, it highlights the importance of attribution in accounting (or failing to account) for empirical phenomena, it provides guidance on architectural development through the explicit statement of the positive heuristic and through the methodological prescription of directing empirical testing at peripheral hypotheses, and it provides a means for evaluating the scientific contribution of an architecture in terms of the success of its positive heuristics.

Acknowledgements

This paper has benefited greatly from extensive and constructive comments provided by John R. Anderson, Christian Schunn and three anonymous reviewers.

References

Altmann, E.M. (2002). Functional decay of memory for tasks. *Psychological Research*, *66*, 287–297.

Altmann, E.M. & Gray, W.D. (2002). Forgetting to remember: The functional relationship of decay and interference. *Psychological Science*, *13*, 27–33.

Altmann, E.M. & Trafton, J.G. (1999). Memory for goals: An architectural perspective. In M. Hahn & S.C. Stones (eds.) *Proceedings of the 21st Annual Conference of the Cognitive Science Society*. pp. 19–24. Mahwah, NJ: Lawrence Erlbaum Associates.

Altmann, E.M. & Trafton, J.G. (2002). Memory for goals: An activation based model. *Cognitive Science*, *26*, 39–83.

Anderson, J.R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Anderson, J.R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, *94*, 192–210.

Anderson, J.R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates. Hillsdale, NJ.

Anderson, J.R. (1991). The place of cognitive architectures in a rational analysis. In K. VanLehn (ed.) *Architectures for intelligence*. pp. 1–24. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J.R. (1993). *Rules of the Mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J.R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*, 313–341.

Anderson, J. R., Albert, M. V., & Fincham, J.M. (2005). Tracing Problem Solving in Real Time: fMRI Analysis of the Subject-Paced Tower of Hanoi. *Journal of Cognitive Neuroscience, 17* 1261–1274.

Anderson, J.R., Bothwell, D., Byrne, M.D., Douglass, S., Lebiere, C. & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *11*, 1036–1060.

Anderson, J.R. & Bower, G. (1973). *Human Associative Memory*. Washington, DC: Winston & Sons.

Anderson, J.R. & Douglass, S. (2001). Tower of Hanoi: Evidence for the cost of goal retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1331–1346.

Anderson, J.R. & Lebiere, C. (1998). *The Atomic Components of Thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General*, *114*, 189–192.

Byrne, M.D. & Anderson, J.R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychological Review*, *108*, 847–869.

Chater, N., & Oaksford, M. (1990). Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition*, *34*, 93–107.

Chong, R.S. (2003). The addition of an activation and decay mechanism to the Soar architecture. In F. Detje, D. Dörner, & H. Schaub (eds.), *Proceedings of the 5th International Conference on Cognitive Modeling*. pp. 45–50. Bamberg, Germany: Universitäts-Verlag.

Chong, R.S. & Laird, J.E. (1997). Identifying dual-task executive process knowledge using EPIC-Soar. In M.G. Shafto & P. Langley (Eds.), *Proceedings of the nineteenth annual conference of the cognitive science society.* pp. 107–112. Mahwah, NJ: Lawrence Erlbaum Associates.

Cooper, R.P. (2006). Cognitive architectures as Lakatosian Research Programs: Two case studies. *Philosophical Psychology*, *19*, 199–220.

Cooper, R.P., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, *85*, 3–44.

Cooper, R.P. & Shallice, T. (1995). Soar and the case for Unified Theories of Cognition. *Cognition*, *55*, 115–149.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–212.

Fodor, J.A. & Pylyshyn, Z.W. (1998). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Goel, V., Pullara, S.D. & Grafman, J. (2001). A computational model of frontal lobe dysfunction: Working memory and the Tower of Hanoi task. *Cognitive Science*, *25*, 287–313.

Hinton, G.E. & Shallice, T. (1989). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Huffman, S.B., Miller, C.S., & Laird, J.E. (1993). Learning from Instruction: A knowledge-level capability within a unified theory of cognition. In *Proceedings of the 15th International Conference of the Cognitive Science Society*. pp 114–119. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hull, C.L. (1943). *Principles of Behavior*. New York, NJ: Appleton-Century.

Hunt, E. & Luce, R.D. (1992). Soar as a world view, not a theory. *Behavioural and Brain Sciences*, *15*, 447–448.

Just, M.A., Carpenter, P.A., & Varma, S. (1999). Computational modelling of high-level cognition and brain function. *Human Brain Mapping, 8*, 128–136.

Kimberg, D.Y. & Farah, M.J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General*, *112*, 411–428.

Kolers, P.A., & Perkins, D.N. (1975). Spatial and ordinal components of form perception and literacy. *Cognitive Psychology*, *7*, 228–267.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (eds.) *Criticism and the Growth of Knowledge*. pp. 91–196. Cambridge, UK: Cambridge University Press.

Laird, J.E., Newell, A. & Rosenbloom, P.S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Laird, J.E. & Rosenbloom, P.S. (1996). The evolution of the Soar cognitive architecture. In D.M. Steier & T.M. Mitchel (eds.) *Mind matters: A tribute to Allen Newell*. pp. 1–50. Mahwah, NJ: Lawrence Erlbaum Associates.

Laudan, L. (1977). *Progress and its problems*. Berkley, CA: University of California Press.

Laudan, L. (1983). The demise of the demarcation problem. In R.S. Cohen & L. Laudan (eds.) *Physics, Philosophy and Psychoanalysis*. pp. 111–127. Dordrecht, The Netherlands: D. Reidel.

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco, CA: Freeman & Co.

McClelland, J.L. & Rumelhart, D.E. (eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2.* MIT Press, Cambridge, MA.

Meyer, D.E. & Kieras, D.E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104*, 3–65.

Meyer, D. E., & Kieras, D. E. (1997b). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review, 104,* 749–791.

Nason, S. & Laird, J.E. (2004). Soar-RL: Integrating reinforcement learning with Soar. In Lovett, M., Schunn, C., Lebiere, C., & Munro, P. (eds.) *Proceedings of the 6th International Conference on Cognitive Modeling.* pp. 208–213. Pittsburgh, PA: Carnegie Mellon University/University of Pittsburgh. Reprinted as Nason, S. & Laird, J.E. (2005). Soar-RL: Integrating reinforcement learning with Soar. *Cognitive Systems Research*, *6*, 51–59.

Neves, D.M., & Anderson, J.R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. pp. 57–84. Hillsdale, NJ: Lawrence Erlbaum Associates.

Newell, A. (1973). You can't play 20 questions with nature and win. In W.G. Chase (ed.) *Visual Information Processing.* pp. 283–308. San Diego, CA: Academic Press.

Newell, A. (1981). The knowledge level. *AI Magazine*, *2*.1–20.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Newell, A. (1992). Soar as a unified theory of cognition: Issues and explanations. *Behavioural and Brain Science*s, *1*5, 464–492.

Newell, A. & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nuxoll, A., Laird, J.E., & James, M.R. (2004). Comprehensive working memory activation in Soar. In Lovett, M., Schunn, C., Lebiere, C., & Munro, P. (eds.) *Proceedings of the 6th International Conference on Cognitive Modeling.* pp. 226–230. Pittsburgh, PA: Carnegie Mellon University/University of Pittsburgh.

O'Reilly, R.C. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press, Cambridge, MA.

Popper, K.R. (1935). *The Logic of Scientific Discovery*. New York, NY: Basic Books. English translation published 1959.

Popper, K.R. (1965). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, UK: Routledge.

Pylyshyn, Z. (1984). *Computation and Cognition.* Cambridge, MA: The MIT Press.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rosenbloom, P.S., Laird, J.E., & Newell, A. (eds.) (1993) *The Soar Papers: Research on Integrated Intelligence*, Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1.* MIT Press, Cambridge, MA.

Seibel, R. (1963). Discrimination reaction time for a 1023-alternative task. *Journal of Experimental Psychology*, *66*, 215–226.

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

Taatgen, N.A. & Anderson, J.R. (2002). Why do children learn to say "Broke"? A model of learning the past tense without feedback. *Cognition*, *86*, 123–155.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Vere, S.A. (1992). A cognitive process shell. *Behavioural and Brain Sciences*, *15*, 460–461.

Wittgenstein, L. (1953). *Philosophical Investigations*. Translated by G.E.M. Anscombe. New York: The MacMillan Company.