Bounding Rational Analysis:

Constraints on Asymptotic Performance

Andrew Howes (HowesA@manchester.ac.uk)

School of Informatics, University of Manchester.

Richard L. Lewis (rickl@umich.edu)

School of Psychology, University of Michigan.

Alonso Vera (avera@arc.nasa.gov)

Carnegie Mellon University and NASA Ames Research Center

Invited paper presented at Integrated Models of Cognitive Systems (IMoCS),

Rensselaer Polytechnic Institute

Workshop & Book (2005).

January 2006

Words = 7887

Abstract

Critical of mechanistic accounts of cognition, Anderson (1990) showed that a demonstration that cognition is optimally adapted to its purpose and environment can offer an explanation for its structure. Simon (1992), in contrast, emphasised that the study of an adaptive system is not a "logical study of optimization," but an empirical study of the conditions that limit the approach to the optimum. In response, we sketch the requirements for an approach to explaining behaviour that emphasises explanations in terms of the optimal behaviour given not only descriptions of the objective and environment *but also* descriptions of the human cognitive architecture and knowledge. A central assumption of the proposal is that a theory explains behaviour if the optimal behaviour predicted by the theory shows substantial correspondence to asymptotic human performance.

1. Introduction

How can we explain human behaviour? Our first purpose in this article is to articulate an approach that emphasises explanations in terms of the optimal behaviour given not only constraints imposed by the objective and environment but also constraints imposed by knowledge and the human cognitive architecture. A second purpose is to describe techniques that realize this approach by supporting the formal reasoning about such constraints. We take as a starting point Simon's (1992) observation that, "behaviour cannot be predicted from optimality criteria without information about the strategies and knowledge agents possess or acquire. The study of behaviour of an adaptive system is not a logical study of optimization but an empirical study of the side conditions that place

limits on the approach to the optimum."… (p. 160). Before articulating our proposal we first review two existing classes of explanation of behaviour: (1) rational explanations of the functions of cognition (rational analysis; Anderson, 1990), and (2) cognitive architecture-based simulations of the mechanisms by which the functions are achieved. These two lines of work provide much of the intellectual framework within which to motivate and understand the present proposal, in particular, understanding both how it is continuous with prior work, and how it departs in significant ways.

Rational analysis

Anderson (1990) emphasised the value of explanations in terms of environment (or at least its experience) and the goals of the cognitive system. He stated a general principle of rationality: "The cognitive system operates at all times to optimize the adaptation of the behaviour of the organism."… (p. 28). Anderson started with the assumption that evolution has to some extent optimized cognition to its environment. He argued that within the limits set by what evolution can achieve, a species is at some stable point in time at a local maximum. Anderson proposed that if the principle of rationality were applied to the development of a theory of cognition then substantial benefits would accrue. In particular, the rational approach (1) offers a way to avoid the identifiability problem, because the theory depends on the structure of an observable world and not on the "unobservable structure in the head"; (2) offers an explanation for why people behave the way they do rather than just how they behave (i.e. because they gain benefit from optimization); and (3) offers guidance to the construction of a theory of the mechanism.

One important implication of these benefits is that rational analysis allows the theorist to avoid the pitfalls of assuming that the human mind is a random collection of mechanisms that are poorly adapted to many of the tasks that people want to achieve. Indeed, explicit in articles advocating rational analysis (including Anderson's) is a critique of mechanistic accounts of cognition on just these grounds. For example, Chater and Oaksford (1999), "From the perspective of traditional cognitive science, the cognitive system can appear to be a rather arbitrary assortment of mechanisms with equally arbitrary limitations. In contrast, rational analysis views cognition as intricately adapted to its environment and to the problems it faces."… (p. 57).

A central and distinguishing feature of rational analysis is that it demands a thorough analysis of the task environments to which cognition is adapted. For example, in light of rational analyses such as Oaksford and Chater (1994), normative accounts of the Wason four card task can be seen to fail precisely because they do not reflect the structure of the general environment experienced by people, but rather are tuned to the simple, unrepresentative tasks studied in the laboratory. At the same time, mechanistic accounts fail when they implicitly adopt the normative analysis and attribute departures from normative behaviour to arbitrary limitations of the underlying cognitive mechanisms. Such considerations led Anderson to critique of his own mechanistic theory of cognition, ACT* (Anderson, 1983) (though see Young and Lewis (1998) for a description of an alternative functional approach to cognitive limitations pursued in the Soar architecture).

Bounding rational analysis

Understanding the relationship between rational analysis and Simon's seminal work on bounded rationality is instructive. Although rational analysis may at first appear at odds with bounded rationality, Anderson (1990) argued that there was no incompatibility between rational analysis and satisficing: it might be rational and optimal to find a normatively satisfactory solution when rational is defined relative to time and processing constraints. Despite this possible in-principle compatibility, Simon (1991, 1992) was critical of rational analysis. His critique focused on the fact that Anderson placed emphasis on the analysis of the environment and backgrounded the role of what Simon called "side conditions". By "side conditions" Simon meant the constraints that were placed on cognition by knowledge, by strategies, and by the human cognitive architecture—the very constraints that rational analysis were intended to abstract away from. Simon (1992), "There is no way to determine a priori, without empirical study of behaviour, what side conditions govern behaviour in different circumstances. Hence, the study of the behaviour of an adaptive system like the human mind is not a logical study of optimization but an empirical study of the side conditions that place limits on the approach to the optimum. Here is where we must look for the invariants of an adaptive system like the mind."… (p. 157).

Accounting for the side conditions: Simulations of behaviour based on cognitive architectures

The most sophisticated current techniques for representing and reasoning about such side conditions are based on computational cognitive architectures (Anderson and Lebier,

1998; Meyer and Kieras, 1997). We focus here on ACT-R, because it uniquely represents the confluence of rational analysis and mechanistic approaches to cognition. Anderson (1990) was clear that there are benefits to both and that the two approaches are complementary. Accordingly, Anderson modified ACT* to reflect the insights gained from rational analyses of memory and choice (Anderson and Milson, 1989; Lovett and Anderson, 1996). The resulting theory, ACT-R (R is for *rational*), combines a model of the decay of activation in declarative memory, derived from the rational analysis of Anderson and Milson (1989), with a model of production rule conflict-resolution derived from a rational analysis of the selection of action on the basis of history of success (Lovett and Anderson, 1996).

One of the strengths of a theory of the cognitive architecture is that it shows how the mechanisms of cognition, perception, and action work together as a single integrated system to produce behaviour. ACT-R has been a spectacular success in just this way. But despite the grounding of ACT-R in rational analysis, models of specific tasks situations constructed in ACT-R are still subject to the rational analysis critique of mechanistic explanations, for two reasons. First, ACT-R integrates a range of mechanisms that are necessary to complete a comprehensive model of human cognition, but that are not directly motivated by rational analysis. Examples include the model of perceptual/motor processing (ACT-R/PM) that was motivated by efforts to build architectures capable of interaction (Byrne and Anderson, 2001; Meyer and Kieras, 1997), and the limits on source activation, which is used to model working memory constraints (Anderson, Reder, Lebiere, 1996). The problem here is not simply that some

components are motivated by rational analyses and others are not –perhaps the more fundamental problem is that motivations of individual components fail to take into account the fact that it is the system as a whole that is adapting to the environment, not components. As cognitive architectures make exceptionally clear, components only have behavioural consequences in conjunction with the set of other components required to produce behaviour.

Second, ACT-R must be provided with specific strategies in the form of production rules in order to perform specific tasks. This is a theoretically necessary feature of architectures (Newell, 1990), but in practice this variable content provides theoretical degrees of freedom to the modeller that may obscure the explanatory role of the architecture in accounting for psychological phenomena. And although Newell's (1990) time-scale analysis was in principle correct –that the architecture shows through at the level of immediate behaviour– recent work is making it increasingly clear that considerable strategic variability is evident even at the level of tasks operating in the sub-second range. A prime example is elementary dual-tasking situations (Meyer and Kieras, 1997), which we turn to below in order to illustrate our new approach.

Local and global adaptation and the role of mechanism and strategy

The current state of affairs can be summarized as follows, and clearly points to new directions for modelling research:

1. A strength of rational analysis is that it provides deeper explanations for both components of the cognitive architecture, and behaviours in particular task situations. These explanations take the form of demonstrations that the components and behaviours are rational adaptations to the structure of the environment viewed from a sufficiently *global* perspective, rather than departures from local optima that point to arbitrary limitations in the underlying cognitive mechanisms.

2. A weakness of rational analysis is that it does not provide a way to systematically and incrementally take into account the side-conditions that bound the approach to optimality in any given *local* task situation. Although it may in fact provide an explanation for some of the side-conditions themselves (to the extent that rational analysis explanations of architectural components are successful), there is no way to systematically draw out the detailed implications of these side conditions for understanding what behaviour is adaptive in a particular task environment.

3. A related weakness of rational analysis is that it does not provide a way to explore the adaptation of the system as a *whole* –and the interaction of all its parts– as opposed than individual components or classes of behaviours.

4. A strength of computational cognitive architectures is that they provide a way to explore the interactions of what Simon called the "side conditions" on the approach to optimality, including knowledge and basic mechanisms of cognition, perception, and action. They can thereby be applied in detail to a wider range of specific task situations

than can rational analysis, and, as ACT-R demonstrates, offer one way to explore architectural mechanisms that may themselves be motivated by rational analysis.

5. A weakness of cognitive architectures is that they do not yield the deep explanations of rational analysis (even if partially grounded in RA-motivated components), because behaviour arises as a function of both architecture and posited strategies, and the latter represents a major source of theoretical degrees of freedom, where strategies may be posited not because they are maximally adaptive, but because they match the empirical results.

In short, we believe that cognitive architectures have made only partial progress in addressing the critique of rational analysis, and rational analysis has made only partial progress in addressing Simon's critique that it has backgrounded the role of mechanistic and knowledge constraints. What we seek is a framework that will permit us to reason about what behaviours are adaptive in a specific local task situation, given a posited set of constraints (architectural and knowledge constraints) on the approach to optimal behaviour, and an explicit payoff function.

The framework we propose here is an initial attempt to achieve this goal. The framework is consistent with rational analysis, but differs from cognitive architectures, in that it values calculation of what is optimal. It differs from rational analysis, and is consistent with cognitive architectures, in that it directly takes into account the complex interaction of architectural mechanisms and how they give rise to the details of behaviour. It differs

from both approaches in that it seeks explanations of specific behaviours as optimal adaptations to both external task constraints and internal system constraints. More precisely, the proposed framework demands that (a) optimality is defined relative to the entire set of constraints acting on the behaving system (internal as well as external constraints), (b) there is an explicit payoff function, known as an *objective function*, and, (c) the optimal performance predicted by the theory corresponds to the empirically asymptotic level of adaptation.

In the discussion section, we briefly consider other related approaches in psychophysics and cognitive modelling. We will also make recommendations for the types of data collection and reporting that is needed for the sorts of analysis that we are proposing. We turn first to a description of the approach and its application to modelling a specific task.

2. How to explain behaviour

In our recent work we have developed an approach designed to complement rational analysis and architectural simulation (Vera, Howes, McCurdy, Lewis, 2004; Howes, Vera, Lewis, McCurdy, 2004). There are three commitments: (1) A commitment to exploring the implications of constraints for the asymptotic bounds on adaptation; (2) a framework for representing theories as sets of constraints; (3) a computational mechanism for calculating the implications of constraints.

2.1. Exploring the Bounds on Adaptation

When people acquire a skill they are able to adapt behaviour so as to incrementally improve the value of some payoff, or objective function. With practice, the scope for improvement attenuates and performance asymptotes. It may asymptote at a level that is consistent with constraints imposed by the environment or perhaps at a level determined by the knowledge that is brought to the task. The bounds may instead be imposed by the human cognitive architecture, including its resource limits (Norman and Bobrow, 1975, 1976). More plausibly, the asymptote may be determined by a combination of constraints, including the stochastic and temporal profiles of the particular task environment and the human cognitive, perceptual, and motor systems.

We assume that under such circumstances people seek to iteratively improve payoff. Effort is oriented towards increasing the value of an objective function that specifies the perceived costs and benefits of action. In seeking to improve a payoff people adapt performance by adopting specific strategies. Improvement eventually asymptotes, and if we ignore for the moment the possibility that people are trapped by local maxima, performance should asymptote at a level where it generates the optimal payoff given the constraints (including those imposed by architecture and knowledge). For a theory of the human cognitive architecture to explain an empirically observed asymptotic bound, substantial correspondence is required between the asymptote and the optimal performance given the theory. A theory that predicts better performance than the observed asymptote is under constrained, a theory that predicts worse performance is over constrained.

Importantly, explanations of the causal role in behaviour of for example, memory, must account for the fact that such component systems have behavioural consequences only when working together with the entire cognitive system. In contrast to rational analysis, the idea is to explore the implications for asymptotic performance of theories of an integrated set of mechanisms applied to a local task environment. The approach can therefore be thought of as a *bounded* rational analysis.

Following Card, Moran, and Newell (1983) we can think in terms of behaviour as being determined by the objective function plus three sets of constraints:

$$\text{Objective} + \text{Task Environment} + \text{Knowledge} + \text{Architecture} \rightarrow \text{Behaviour}$$

Each set of constraints generally *underspecifies* behaviour in the absence of an explicit objective function. Thus, the task environment alone affords a large space of possible behaviours; this space is further constrained by architecture constraints, For our present purposes, this space of possible behaviours represents the space of possible *strategic variations* and may be constrained yet further by knowledge constraints. The objective function then selects a single surface in this subspace that represents the optimal set of possible behaviours satisfying the joint set of constraints.

Obviously, an explanation in terms of objective and environment (i.e. a rational analysis) is to be preferred to a bounded rational analysis on the grounds of parsimony. Such

explanations are possible when the objective and the environment constraints alone yield a subspace of behaviours that corresponds with observed behaviour. However, in many circumstances such explanations are not possible.

2.2. A framework for representing theories as sets of constraints

The second requirement is for a theoretical ontology that provides a language for expressing constraints on information processing mechanisms. By definition, information processes receive, transform, and transmit information. A process receives and transmits information from and to other processes. We assume that in order for two processes to exchange information they must overlap in time, or each must overlap in time with a common mediating, or buffering process, which stores the information for some, perhaps very short, period of time. McClelland (1979) introduced the hypothesis that information processes were cascaded, that is that they overlapped in time, and in addition that the quality of information passed from one process to another increased with time.

Our version of cascade theory commits to the following assumptions: (1) Processes must overlap in time if they are to transfer information; (2) a process is executed by a processor (also known as a resource); (3) some function relates the accuracy of information produced to the duration since the process started (Howes, et al., 2004). (It follows from 3 that a process has a minimum duration, before which no transmission occurs, and a maximum duration, after which no transmission occurs.)

In addition to a framework for representing the constraints on information flow we need some commitment to the process and processing capabilities of the human cognitive architecture. What processes characterize human cognition? What kinds of processors are they executed by? Here we are interested in an account in which information processing is conceived of in terms of an interacting set of processes each with defined resource requirements, temporal duration, and input/output characteristics. As a starting point we take Card, Moran, and Newell's (1983) Model-Human Processor (MHP). The representation of processes abstracts over representation and algorithm. For the purposes of explaining behaviour it is not always necessary to define the precise mapping between input and output representations of individual processes. It is sufficient for example to state that a stimulus is perceived and that a response is retrieved in some mean time with some standard deviation.

2.3. Calculating the implications of constraints on behaviour

The third requirement is for a language in which constraints on human information processing (Section 2.2) can be specified in a computable form and the implications for the asymptotic bound on performance calculated.

The fact that human performance depends on a multiplicity of complex interacting constraints deriving from the environment, from the human cognitive architecture, and from knowledge makes calculating the implications of constraints difficult. Skilled performance of a routine task usually involves the execution of a number of parallel but interdependent streams of activity: For example, one hand may move to a mouse; while

the other finishes typing a word; and the eyes begin to fixate on a menu while the required menu label is retrieved from memory. Each of these processes takes a few hundred milliseconds, but together they form behaviours that take many seconds. Importantly, the details of how processes are scheduled has significant consequences for the overall time and resource requirements.

Vera et al. (2004) proposed that one response is to represent theories as sets of constraints using a predicate calculus constraint logic. A constraint is simply a logical relation between variables. Constraint satisfaction has the potential to provide a formal framework for the specification of theories of interactive cognition, and thereby for the construction of mathematically rigorous tools for supporting the prediction of the bounds that the constraints imply for adaptation. Of central importance is the fact that constraints are declarative and additive. They are declarative in that relationships between variables can be stated in the absence of a mechanism for computing the relationship. They are additive in the sense that the order in which constraints are specified does not matter. These properties should allow theoretical assumptions to be expressed in a computable form that is relatively independent of the arbitrary constraints that are sometimes imposed by the machine, or software algorithms, with which computation is conducted.

The fact that constraints allow the specification of what is to be computed without specification of how the computation is carried out (the algorithm), means that considerable flexibility is enabled in the desired properties of the schedule. Importantly, it does not matter which algorithm is used to derive the optimal solution (as long as it

works!). It happens that our previous work has made use of a branch-and-bound algorithm (Howes et al., 2004; Vera et al., 2004) but tools based on dynamic programming or whatever other algorithm would do just as well. Similarly, Monte Carlo simulation can be used to generate an estimate of the optimal adaptation as long as care is taken to search the space of possible strategies within which the optimal solution in located.

The point of our paper is not to argue for the value of a particular optimisation algorithm but to argue for the scientific utility of considering the relationship between the optimum under constraints and the empirically observed asymptote. From the perspective of the scientific aims, the set of possible optimal solutions is defined precisely by the payoff function (the objective) and by the set of (declarative) constraints: There is no need to specify the optimization algorithm in order to specify a theory; the optimization algorithm is simply the means by which one derives the implications of the theory.

3. Example: Constraints on dual task performance

Consider how we might predict performance on simple Psychological Refractory Period (PDP) tasks. For example, in Schumacher et al.'s (1999) experiment 3 participants were required to respond to a tone and a visual pattern (simple or complex) with key presses that depended on whether the tone was high or low and whether the pattern contained a particular feature. The tone and the pattern were presented with a small gap of between 50 and 1000ms (Stimulus Onset Asynchrony - SOA). Participants were asked to prioritize the tone task (task 1) over the pattern task (task 2). The tone task response

times were, on average, unaffected by SOA. In contrast, the mean pattern task response time, at a short SOA (50ms), was less than the sum of the tone task and pattern response times at long SOAs (> 500ms). This finding has been taken as evidence that some elements of tone and pattern task were performed in parallel at short SOAs. Byrne and Anderson were interested in modelling Schumacher's data using ACT-R/PM in order to demonstrate that cognitive parallelism is not required to explain these results. They argued that the results can be modeled with either the EPIC or ACT-R/PM assumptions and that Schumacher's data provides evidence for strategic deferment of the pattern task response.

3.1. Optimizing over the statistics of interaction

Our previous work, Howes et al. (2004), demonstrated the potential analytic role of optimization (as described in Section 2 of this chapter) in exploring the space of possible adaptations. What it did not do was articulate how constraint analyses can be used to explore how people adapt to the statistics of interaction with an uncertain environment and with uncertainty in the duration of internal processes. One such adaptation is required in the PRP task where participants need to ensure the ordering of task 1 and task 2 responses despite fluctuations in the durations of each response. Here we develop a constraint model of a generic PRP task and describe its predictions.

Asked to respond as quickly as possible to a single stimulus an individual will produce a range of approximately normally distributed responses. In a dual task situation, such as the psychological refractory period task, each response has its own distribution. If in the

dual task situation there is some benefit, a gain, from responding rapidly and some cost to making a response reversal, responding to task 2 (the pattern task) before task 1 (the tone task), then participants will weigh the costs and benefits of fast and slow responses. Parameterised with the response means for individual responses, their standard deviations, and estimates of the costs and benefits of the space of possible behaviours, an adequate theory of the human cognitive architecture must predict the asymptotic mean and standard deviations of the response times in a dual task scenario.

Imagine an architecture A' in which there are no mechanisms by which the processing of task 1 can influence the processing of task 2. That is there are no necessary task interactions and no shared cognitive or perceptual/motor resources. How do we test whether A' can predict and explain performance on a PRP task? According to the assumptions of our framework we need to determine whether the best possible performance predicted by the architecture corresponds to the asymptotic human performance. If the processing for each task is entirely independent then the extent to which the response distributions overlap will determine the frequency with which response reversals occur. If participants intend to avoid response reversals then a strategy is required. In the case of the very simple architecture A' the only strategy available that mitigates against response reversals is to delay the response to task 2. By delay we mean to wait a fixed, trial independent, amount of time that is added to what would otherwise be required to make the task 2 response. This simple strategy will temporally separate the task 1 and task 2 response distributions.

If the best available strategy is to delay the task 2 response then the next question is by how much? If we know what the payoff function is for participants (i.e. how much they gain for a correct response and how much they lose for a response reversal) then we can derive exactly the optimal amount of time to delay response 2. According to the theory, a participant should select a value to delay task 2 that is consistent with the constraints and which maximizes the value of the payoff.

We can specify the constraints and the objective function (the payoff) for different values of the delay as a constraint model. However, rather than fixed durations, here we sample duration from a normal probability distribution.

Constraints

SOA in { SOAmin... SOAmax }

$RT1_i$ = normal(M1, SD1)

$RT2_i$ = normal(M2, SD2) + DELAY

In addition to the constraints, the objective captures a speed/accuracy trade-off between going fast and avoiding response reversals. The payoff for a trial is the gain minus the total time cost and minus the cost of reversal. The time cost is defined as a weight times the duration of the latest of the two responses. The cost of reversal is defined as some weight times 1 if a response reversal occurred and 0 otherwise. Higher values of the strategically set delay variable will tend to decrease the proportion of response reversals but at the cost of increasing the total time required to perform the task.

Payoff: For trials 1 to N,

Average Payoff = ($\sum i=1$ to N : GAIN – C_t - C_r) / N - (1)

Time cost $C_t = W_n$ x max($RT1_i$, (SOA + $RT2_i$)))

Cost of reversal $C_r = W_m$ x f(SOA + $RT2_i$ – $RT1_i$))

f(X > 0) = 0

f(X =< 0) = 1

Despite the simplicity of this model, to our knowledge none of the reported PRP data sets are suitable for testing its validity. While some experiments, such as those reported by Schumacher et al. (1999), were controlled for the cost/benefit trade-off between a fast response and a response reversal, neither the payoff achieved by participants, nor details of standard deviations and reversals rates are reported.

The constraint model makes predictions dual task performance given parameters determined from single task performance. What needs to be calculated is a prediction of the asymptotic performance time and error rate at short SOAs given (a) mean and standard deviations of performance time at long SOAs and (b) an experimental paradigm in which participants are subject to a payoff regime enforcing a trade-off between RT and rate of response reversal. Importantly, in order to test the model, each trial would take a fixed amount of time independently of the response time. This assumption eliminates the

additional benefit of a rapid response, beyond the reward that determines the weightings in the objective function, though other payoff regimes are possible.

Calculation of the optimal value of the task 2 delay requires Monte Carlo simulations for each potential duration. The payoff achieved on each trial can then be aggregated to give a total payoff for each possible strategy (value of the delay). For A' there will be an N-shaped relationship between duration of delay and payoff. The optimal performance (maximal payoff) implied by A', and therefore the predicted asymptote on human performance, given task environment, constraints, and payoff function will correspond to the peak of this curve.

Note that unlike in model fitting methodologies (e.g. as used in Meyer and Kieras (1997) to determine the length of the defer process) we are not proposing to choose a value of the delay parameter so that the model fits the data. Rather we are claiming that the model predicts that participants will delay task 2 by a particular duration (the optimal duration of the delay process), subject to an estimable confidence interval, and given only parameters set from single task performance. If at asymptote participants delay by more or less than the predicted duration then the A' theory is wrong or at least incomplete. To the extent that the theory cannot be successfully modified by adding or removing constraints, we have learned something interesting about the limits on the mechanisms of adaptation.

3.2. Discussion

The example that we have developed above demonstrates a method for calculating the asymptotic bound predicted by a theory. However, the constraints are based on overly simplistic assumptions about the internal processing mechanisms. Calculating the implications of more elaborate theories such as Byrne and Anderson's (2001) or Meyer and Kieras's (1997) over the statistics of interactive behaviour requires constraint models such as those developed in Howes et al. (2004).

Comparison of the predicted asymptote to the observed asymptote provides a test of the adequacy of the theory. If people do not perform as well as the predicted asymptote then the implication is that the theory is under-constrained. If people perform better than the predicted asymptote then the implication is that the theory is over-constrained. It follows that there is no role for the notion of suboptimality within the explanatory framework that we have described. If global maxima are discoverable and if optimality is defined relative to the entire set of constraints and the objective function rather than relative to the task and environment only then the extent to which the optimal performance corresponds to the asymptotic human behaviour is a measure of the goodness of the theory, not of the suboptimality of the human behaviour.

4. General Discussion

We have sketched the requirements for an approach to explaining behaviour that emphasises the importance of explanations in terms of the optimal behaviour given not only descriptions of the objective and environment *but also* descriptions of the human

cognitive architecture and knowledge. We illustrated the approach with a model of strategic processing in Psychological Refractory Period (PRP) tasks. The model made limited assumptions about response variance and we described how a prediction of dual-task response separation given an objective function that traded time taken against response reversal could be derived. We claimed that a theory could be said to explain the data if it could be established that there was substantial correspondence between the optimal performance implied by the theory and the asymptotic performance observed in human behaviour. In the remainder of the general discussion we (a) describe related work; (b) describe how to design experiments that provide data amenable to constraint-based explanations; (c) reflect further on how a bounded rational analysis complements rational analysis (Anderson, 1990).

Related Work

The approach to explaining cognition that we have discussed was motivated in part by Roberts and Pashler (2000) and also by Kieras and Meyer (2000). Both have observed the potential problems with failing to explore the contribution of strategies and architectural constraints to the range of possible models of human performance. Kieras and Meyer (2000) responded by proposing the use of a *bracketing heuristic*. A bracket was defined by the speed of the fastest-possible strategy for the task, and the slowest-reasonable strategy. Kieras and Meyer predicted that observed performance should fall somewhere between the performance of these two strategies. They also articulated the

importance of exploring the space of strategies to explaining the phenomena being modeled. While there are similarities, there are two differences to our approach. First Kieras and Meyer (2000) focused on bracketing the *speed* of strategies rather than their payoff. Second, they saw bracketing as a means of coping with the problem that optimisations cannot be forecast. Kieras et al. state that bracketing was a way to construct, "truly predictive models in complex task domains where the optional strategy optimizations users would devise cannot be forecast."... (p. 131).

Others have also used analyses of optimal performance to bracket predictions. There is a long and active tradition in analyses of optimal performance in psychophysics (Swets, Tanner, Birdsall, 1961; Trommershäuser, Maloney and Landy, 2003; Geisler, 2003). More recently authors such as Kieras and Meyer, (2000) and Neth, Sims, Veksler and Gray (2004) have contrasted human performance to optimal performance on more complex cognitive tasks. Neth et al. (2004) used analysis of the best possible performance given a particular strategy to predict a bracket for behaviour on a decision making task. Others, for example Fu and Gray (2004) and O'Hara and Payne (1998), have exposed apparent suboptimalities in behaviour and sometimes offered explanations that allow those behaviours to be interpreted as rational adaptations given additional constraints.

Exploring optimality criteria has been particularly fruitful in psychophysics. An Ideal Observer Theory is a computational theory of how to perform a perceptual/cognitive task

optimally given properties of the environment and the costs/benefits associated with different outcomes (Geisler and Diehl, 2003). However, Geisler and Diehl (2003) state: "While ideal observer theory provides an appropriate benchmark for evaluating perceptual and cognitive systems, it will not, in general, accurately predict the design and performance of real systems, which are limited by a number of factors..." Also, Geisler (2003): "Organisms generally do not perform optimally, and hence one should not think of an ideal observer as a potentially realistic model of the actual performance of the organism. " Geisler and Diehl (2003) particularly focus on the fact that the real observer may correspond to a local maximum in the space of possible solutions, whereas the ideal observer corresponds to the global maximum. The Ideal Observer corresponds to Marr's computational theory and is a theory of what the organism should compute given the task and stimuli (Geisler and Diehl, 2003), not what it is rational to compute given the entire set of constraints.

Kieras and Meyer (2000) and Geisler and Diehl (2003) may be right, in general, to be pessimistic about the prevalence of task domains in which it is possible to forecast people's strategy optimizations. However, if task domains where it is possible can be identified and if it is accepted that architectures show through at the limit, particularly when resources are limited (Norman and Bobrow, 1975, 1976), then these task domains may be particularly useful for evaluating theories of the human cognitive architecture: The optimal solution given the theory can be taken as a forecast of the asymptote. In addition, Kieras and Meyer's view may have been influenced by the lack of available

techniques for calculating the optimal solution given a complex and heterogeneous set of constraints on information processing models. Our previous work (Howes, Vera, Lewis, McCurdy, 2004; Vera, Howes, McCurdy, Lewis, 2004) has articulated general purpose analytic techniques for predicting strategy optimizations.

Experimental Methodology

The predictions made by the analyses could not be tested against results from experiments that failed to control for the trade-off between speed and accuracy. Unfortunately, despite the work of Meyer and Kieras (1997) the absence of controls on speed/accuracy trade-offs is wide spread in experimental cognitive psychology. With a few exceptions, error rates tend to be dismissed as small or are excluded from analysis presumably motivated by the view that they are an aberration, just noise that distracts from the main picture. The reality is that human adaptation to the objective function within the limits set by the constraints is pervasive. People adapt enthusiastically and continuously (Charman and Howes, 2003). They adapt tasks that take hundreds of seconds to complete and they adapt tasks that take hundreds of milliseconds to complete. To understand this adaptation it is critical to fully understand the objective to which participants are adapting.

Knowing what the participant was instructed is probably not sufficient (Kieras and Meyer, 2000). Participants do not merely do what they are told to do, rather they

interpret instructions to generate objective functions that are consistent with longer-term traits. A challenge is to find experimental paradigms for resource-limited tasks that expose participant's objective functions and thereby support the rigorous calculation of the predicted asymptote. While the work of Trommershäuser, Maloney and Landy, (2003) provides an example of what can be done for pointing tasks, more needs to be done for tasks that involve sequential ordering.

Bounding Rational Analysis

The role of optimality criteria in cognitive science has been controversial. Indeed, one objection to our approach might be: People do not optimise, they satisfice. The critique would seem consistent with Simon's critique of the assumption that people make optimal economic decisions (Simon, 1957). But this would be to miss the fundamental distinction between the idea that Simon rejected (i.e. the hypothesis that people are optimally adapted to the environment) and the idea proposed in this article: That given the adaptive nature of human cognition, an explanation of behaviour must explain why people do not do better than they do. It must explain the approach to the asymptote in terms of the implications of psychological bounds. Despite a shared value in determining optimal adaptations, the approach that we have described is not rational analysis. Where Anderson emphasised optimality in terms of the task and environment, the approach that we have articulated gives equal emphasis to constraints on architecture and knowledge. Our approach is more closely aligned with Simon (1992) who emphasised the need to

investigate the side conditions that place limits on the approach to the optimum.

In fact, to the extent that predictions made through optimisation are constrained by hypotheses concerning internal processing limits, the predictions are not optimal relative to the goal and environment. Our approach is therefore consistent with Simon's reminder that explaining behaviour requires reference to internal processing limits and capabilities and it is consistent with the idea that people satisfice. The challenge that we are addressing could be characterised, perhaps, as how to precisely articulate what it means to satisfice.

We also expect that there are many task environments where incremental improvement is unlikely to lead to an optimal solution. In these cases suboptimal performance may result from too many local maxima. Here incremental improvement may lead to an asymptote but not the asymptote that corresponds to the optimal solution given the theory.

In general, task environments where the optimal solution is within the grasp of incremental improvement may be more suitable for evaluating the consequences of theories of psychological resources for the asymptotic bound on the adaptation of behaviour. For these environments, the absence of a correspondence between the optimum implied by a theory and the behavioural asymptote is evidence for the inadequacy of the theory. In contrast, in task environments where the optimal solution is unlikely to be generated incrementally, the absence of correspondence could be due to

either the shape (availability of local maxima) of the task environment or to the inadequacy of the theory. Behaviour in these task environments is unlikely to offer a good basis for empirical tests of a theory of what bounds adaptation.

Analyses similar to that which we have described in this paper could assist the development of rigorous answers to questions of optimality and therefore rationality in interactive cognitive skill. Questions have been raised by a number of authors about the extent to which people make optimal adaptations (Fu and Gray, 2004; Gray and Boehm-Davis, 2000; Taatgen, 2005). First, to explain cognition, optimality must take into account constraints on architecture and knowledge, not only constraints on the environment. Second, a claim that behaviour is suboptimal or biased, or that it is not rational, does not explain behaviour. A claim of suboptimality carries little content in the absence of an explicit theory of what *is* optimal, and a means for calculating the implications of that theory. An apparent suboptimality raises the question as to what modification is required to the theory so as to align the predicted performance bound with the empirically observed asymptotic bound.

Conclusion

To conclude, we have argued that neither rational analysis nor computational simulation are sufficient approaches to explaining cognition. Another promising approach may be to test for correspondence between theories of optimal performance given both

environmental and psychological constraints, and empirically observed asymptotic bounds in particular task environments. Such an approach requires not only theories of the constraints imposed by the task environment but also theories of the constraints imposed by the cognitive architecture and by knowledge, it not only requires exploration of the trajectories through the space of possible adaptations but also systematic analysis of the bounds on that space.

5. References

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

Anderson, J.R. (1990). *Rational Analysis*. Mahwah, NJ: Erlbaum.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology, 30*, 221-256.

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

Anderson, J.R. & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review. 96* (4), 703-719.

Byrne, M.D. & Anderson, J.R. (2001). Serial modules in parallel: The psychological refractory period and perfect time sharing. *Psychological Review, 108* (4), 847-869.

Card, S.K., Moran, T.P., Newell, A. (1983). *The Psychology of Human Computer Interaction*. NJ: Erlbaum.

Charman, S.C. & Howes, A. (2003). The adaptive user: an investigation into the cognitive and task constraints on the generation of new methods. *Journal of Experimental Psychology: Applied, 9*, 236-248.

Chater, N. & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science, 3*(2), 57-65.

Fu, W.-T., & Gray, W.D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science, 28*, 901-935.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825-837). Boston: MIT Press.

Geisler, W.S. & Diehl, R.L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science 27*(3) 379-402.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds Matter: An introduction to microstrategies and to their use in describing and predicting interactive behaviour. *Journal of Experiment Psychology: Applied, 6*(4), 322-335.

Howes, A., Vera, A. H., Lewis, R. L., & McCurdy, M. (2004). Cognitive constraint modelling: A formal approach to supporting reasoning about behaviour. In K. D. Forbus & D. Gentner & T. Regier (Eds.), *26th Annual Meeting of the Cognitive Science Society, CogSci2004* (pp. 595-600). Hillsdale, NJ: Lawrence Erlbaum Publisher.

Kieras, D. E., & Meyer, D. E. (2000). The Role of Cognitive Task Analysis in the Application of Predictive Models of Human Performance. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis* (pp. 237-260). Mahwah, NJ: Erlbaum.

Lovett, M. and Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology, 31*, 168-217.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287-330.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive

processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, 104, 3-65.

Neth, H., Sims, C. R., Veksler, V. & Gray, W. D. (2004). You can't play straight tracs and win: memory updates in a dynamic task environment. In K. D. Forbus, D. Gentner & T. Regier (Eds.). *Proceedings of the Twenty-Sixth Annual Meeting of the Cognitive Science Society* (pp. 1017-1022). Hillsdale, NJ: Lawrence Erlbaum.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*, 44–64.

Norman, D. A., & Bobrow, D. G. (1976). On the analysis of performance operating characteristics. *Psychological Review, 83*(6), 508–510.

O'Hara, K.P, & Payne, S.J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology, 35*, 34-70.

Oaksford, M. & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103*, 381-391.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Ruthruff, E., Johnston, J.C., Van Selst, M., Whitsell, S., & Remington, R. (2003). Vanishing dual-task interference after practice: Has the bottleneck been eliminated or is it merely latent. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 280-289.

Schumacher, E. H., Lauber, E. J., Glass, J. M., Zurbriggen, E. L., Gmeindl, L., Kieras, D. E., & Meyer, D. E. (1999). Concurrent response-selection processes in dual-task performance: evidence for adaptive executive control of task scheduling. *Journal of Experimental Psychology: Human Perception and Performance , 25*, 791-814.

Simon, H.A. (1957). *Models of Man*. New York: Wiley.

Simon, H.A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (ed.) *Architectures for Intelligence: The 22nd Carnegie Mellon Symposium on Cognition*. Hillsdale, NJ: Erlbaum.

Simon, H.A. (1992). What is an "explanation" of behaviour? *Psychological Science, 3,* 150-161.

Swets, J.A., Tanner, W.P. Jr, & Birdsall, T.G. (1961). Decision processes in perception. *Psychological Review, 68*, 301-40.

Taatgen, N.A. (2005). modelling parallelization and speed improvement in skill acquisition: from dual tasks to complex dynamic skills. *Cognitive Science, 29*, 421-455.

Trommershäuser, J., Maloney, L. T. & Landy, M. S. (2003). Statistical decision theory and tradeoffs in motor response. *Spatial Vision, 16*, 255-275.

Vera, A., Howes, A., McCurdy, M., and Lewis, R.L. (2004). A constraint satisfaction approach to predicting skilled interactive cognition. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 121-128). New York, NY: ACM Press.

Young, R.M. and Lewis, R.L. (1998). The Soar Cognitive Architecture and Human Working Memory. In Miyake, A. and Shah, P. (Eds), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York: Cambridge University Press.