Made to Measure: Ecological Rationality in Structured Environments

Seth Bullock and Peter M. Todd
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, D-14195 Berlin, Germany
Email: bullock@mpib-berlin.mpg.de

July 23, 1999

Abstract

A working assumption that processes of natural and cultural evolution have tailored the mind to fit the demands and structure of its environment begs the question: how are we to characterize the structure of cognitive environments? Decision problems faced by real organisms are not like simple multiple-choice examination papers. For example, some individual problems may occur much more frequently than others, whilst some may carry much more weight than others. Such considerations are not taken into account when (i) the performance of candidate cognitive mechanisms is assessed by employing a simple accuracy metric that is insensitive to the structure of the decision-maker's environment, and (ii) reason is defined as the adherence to internalist prescriptions of classical rationality. Here we explore the impact of frequency and significance structure on the performance of a range of candidate decision-making mechanisms. We show that the character of this impact is complex, since structured environments demand that decision-makers trade off general performance against performance on important subsets of test items. As a result, environment structure obviates internalist criteria of rationality. Failing to appreciate the role of environment structure in shaping cognition can lead to mischaracterising adaptive behavior as irrational.

Keywords: decision making, frequency structure, significance structure, rational synthesis, adaptive behavior, externalism.

Running head: Rationality in Structured Environments

1 Introduction to the problem of environment structure

Organisms are matched to the demands of particular environments. Deep-sea creatures, for instance, have evolved to require a high pressure aqueous environment, and to exploit the opportunities that this environment affords (such as profound darkness as a backdrop for bioluminescence) in order to effect their survival and reproduction. When taken out of the environment that they are adapted to, such creatures can suffer explosive consequences. Within biology this vital match between a biological system and its environment is termed "fit". The environment to which an organism is fitted by evolution is known as its "niche".

In much the same way that biological devices are matched to their niches, decision-making mechanisms are also matched to particular kinds of task (see, e.g., Gigerenzer, Todd, & the ABC Group, 1999). As in the case of biological fit, the suitability of these cognitive mechanisms is predicated on the structure of their environment. The success of a particular cognitive mechanism will depend not only upon the task demanded of it, but also the nature of the problem it faces in achieving this task. Whilst a tin-opener is suited to the task of opening tins, it may not be suited to particular tins (such as oil drums, etc.) — its limitations make it ill-fitted to certain problems but suitable for others. The extent to which an organism fits its niche, or a mechanism matches the problem it faces, is the extent to which it meets the demands of its environment.

These considerations imply straightforwardly, that different environment structures will, by definition, favor different cognitive mechanisms. Thus, to evaluate the performance of these mechanisms, we have to take environment structure into account. But what is environment structure and how are we to measure it? Here we concentrate on the ramifications of two well-specified aspects of environment structure on the performance of cognitive decision-making mechanisms.

To appreciate these two forms of environment structure, imagine that you are a university professor. Every once in a while, a student who has been offered a similar job by two universities approaches you for your advice. Which job offer should they accept?

1. Since neither job applications nor offers of employment are made at random, one might expect certain universities to feature more frequently than others in this kind of decision.

2. Since not all universities have equal status, some decisions of this kind may be more significant than others.

Suppose that your students know that across all the possible pairs of universities, your advice is correct 80% of the time. Suppose that they also know that a colleague of yours is only correct 70% of the time. Should they approach you for advice rather than your less knowledgeable colleague? Not necessarily. Despite the higher accuracy of your advice across *possible* problems, the students may quite rightly reject you if the 20% of cases in which you err are the most *important* or *frequent* ones, while your colleague does not make these frequent, costly mistakes, but rather errs only in trivial or uncommon circumstances.

Notice that in this example, general-purpose knowledge (high accuracy across possible test items) has been sacrificed for special purpose knowledge (high accuracy across frequent or significant test items). Notice also that failure to appreciate either frequency or significance structure in this example will lead observers to conclude that students are acting irrationally in choosing the less knowledgeable professor.

Putting ourselves in the shoes of the job-seeking student, how should we assess the performance of each professor before deciding whose advice to heed? We might carefully select specific test items which we expect to best discriminate between hypotheses regarding the professors. Whilst patterns of success and failure across such a set of diagnostic test items may reveal facts about how the professors go about solving their task, the performance over such a set will not be representative of the professors' performance in general unless this set of test items is itself representative.

Similarly, assessing the performance of each professor using a multiple-choice paradigm in which (i) the answer to each test item is weighted equally, and (ii) either every possible test item is presented once, or a uniform random sample of possible test items is presented, will also fail to capture the underlying structure of the problem, and therefore will misjudge any decision-making mechanism adapted to that structure.

Assessing each professor on a representative or "natural" sample (Brunswik, 1955) of test items is the only way to reasonably decide between them. This approach to assessment and the role of environmental considerations derives from an ecological perspective on rationality which itself follows from the evolutionary biology considerations with which this paper opened. In the next section we present the foundations of this notion of ecological rationality, before turning to specific examples of frequency structure and significance structure, and their implications for decision making in structured environments.

2 Ecological Rationality

Consider two contrasting assumptions about how best to conceive of cognitive mechanisms. The first stems from an observation about origins.

- Assumption: Processes of natural and cultural evolution (sometimes via the lifetime learning fashioned by these processes) have tailored the mind to fit the demands and structure of its environment. Behavior must be adaptive, i.e., suited to its proper environment, to be successful.

This assumption invokes a natural process (evolution) and an externalist criterion of success (the environment). It has a direct implication.

- Implication: The assessment of candidate cognitive mechanisms must be sensitive to facts concerning environment structure.

The second conception of cognitive mechanisms considers them to approximate general-purpose, optimal (and ultimately mythical) devices. It is thus an assumption about goals.

- Assumption: Minds are best understood as approximating a Laplacean superintelligence (Laplace, 1951), which will, by definition, achieve general-purpose, optimal performance in any situation, no matter how rare; for any price, no matter how costly; and for any reward, no matter how meager.

This assumption invokes an ideal, and implies internalist criteria for success.

- Implication: General purpose performance cannot, by definition, rely upon assumptions about the problem to be faced, hence the behavior of candidate cognitive mechanisms should conform to internalist rational criteria, e.g., coherence, transitivity, etc., since it is through the adoption of these criteria that a superintelligence will achieve its optimal performance.

Whilst this second, classically rational approach to cognition is somewhat of a straw man, the internalist criteria which it promotes are widespread within decision-making psychology and related fields, taking the form of prescriptive norms; Your Subjective Probabilities Must Sum to Unity! Be Transitive in Your Choices! Be Coherent! Be Consistent in Your Preferences! In contrast, the first approach to cognition embraces an ecological perspective on rationality, dispensing with internalist criteria in favor of an externalist performance metric. In the same way in which evolutionary biology assesses the fitness of adaptations in terms of the extent to which they perform the task for which they were selected, ecologically rational reasoning is reasonable to the extent that it is successful within its proper environment.

The perspective on cognition afforded by the concept of ecological rationality is a powerful one. Understanding its rationale requires that certain lay terms be given technical meanings. Although space limitations prevent a full account of its derivation, a few of the more pressing issues will be briefly addressed here (readers are directed to Millikan, 1984, for an account of the role of evolution in underwriting the attribution of functions to cognitive mechanisms).

2.1 Proximality and Proxihood

First, the phrase "proper environment" is used here (e.g. in the first assumption above) in the same technical sense in which Millikan (1984, 1993) employs the term "Normal conditions" to mean "the conditions to which [a] device ... is biologically adapted" (Millikan, 1984, p.34). This biological adaptation is ultimately evolutionary, but may also involve learning, as in the case of a mechanism which has evolved to detect mates, but is calibrated through some period of juvenile experience. The nature of a mechanism's proper environment must typically be established historically since it will usually be a past environment, although as noted above, a mechanism which is calibrated by individual learning of some kind may be properly suited to its current environment, or at least to the environment in which it was calibrated. In general, the proper environment cannot be established statistically by establishing what the current environment of a mechanism typically is.

Since our knowledge of past environments will generally be poor, establishing the structure of these environments with the degree of precision necessary in order to predict, in one fell swoop, the adaptations which resulted from them may be hard, if not impossible. However, taking as a working assumption the hypothesis that, whatever these environments were, they have shaped the character of extant cognitive mechanisms allows us to approach cognition and behavior as evidence from which to infer the adaptive tasks faced by our ancestors and the structure of the past environments in which our ancestors had to achieve them (c.f. the evolutionary psychology approach to studying evolved cognitive mechanisms as laid out by Cosmides & Tooby, 1987). This approach is clearly circular: current behavior is used to infer past environments which are in turn used to predict current behavior-generating mechanisms. However, this circularity is not vicious. Each turn of the cycle produces new behavioral hypotheses which can be tested and used to revise our environmental assumptions. This process is analogous to that employed by the proponents of rational analysis (Anderson, 1991) who iterate through a similar cycle, repeatedly revising the nature of a decision problem until the optimal solution to this problem matches the observed performance of the natural decision makers they are interested in.

Second, in using terms such as "success" and "task" when describing the performance of a natural mechanism, we are eliding an important dimension. The manner in which these terms should be interpreted depends on whether one is concerned with explanations which are biologically *ultimate* or more *proximate*. Whilst ultimately every biological adaptation has been selected for the task of effecting its own reproduction, with appropriate caveats, organisms and the organs they contain can also be considered to face more proximate adaptive subgoals (e.g., pumping blood, regulating body temperature, finding food, seducing a mate). Similarly, although the success of a natural cognitive mechanism is ultimately cashed out in the same fitness terms as any biological adaptation, its performance can be understood more proximately in terms of its reasoning success. This reasoning success can be considered as a proxy for the biological fitness of a reasoning mechanism.

However, establishing a proxihood relationship between some measure of successful reasoning and ultimate fitness is not straightforward. For example, the capture of accurate information is often considered to be a good measure of reasoning success (e.g., Oaksford & Chater, 1994, 1996, but see also Klauer, 1999). In his model of animal communication, Grafen (1990) equates the success of a choosy peahen with her accuracy in capturing the mate value of her suitors. One might expect that to the extent that a reasoning mechanism tends to provide veridical information to the deliberation or action systems which depend on this information, such a reasoning mechanism would be fit. However, using the capture of veridical information as a proxy for fitness ignores the possibility that even accurate information may sometimes be epistemically worthless (Evans & Over, 1996).

For example, a decision-making mechanism used by a peahen to judge the quality of peacocks may provide equally accurate assessments in two cases, yet if the first case involves a poor quality suitor and the second a high quality suitor, the value of these two pieces of equally accurate information will differ greatly. The first assessment allows the peahen to confidently reject a poor suitor, avoiding the costly mistake of making a long-term investment with a poor-quality mate. In contrast, the second assessment allows her to confidently accept a good suitor, avoiding the (presumably) much less costly mistake of overlooking the currently available good mate. Thus, these two decisions have radically different implications for the peahen's fitness and hence the fitness of the peacock-assessing mechanism that she employs. Moreover, for species in which both sexes are choosy, whether a female's assessment of a particular potential mate is accurate or not may have no impact on her fitness if the suitor being assessed rejects her (Todd & Miller, 1999).

These examples highlight the fact that it is the behavior which results from an organism's reasoning rather than the reasoning itself which is the locus of selective pressure. Whilst accurate and error-free reasoning is clearly typically a conduit leading to adaptive behavior, it does not follow that "irrational" reasoning must have negative consequences for the success of an organism's behavior. As we move along the explanatory dimension from explanations of decision-making behavior in terms of some ultimate goal (reproduction) to explanations in terms of increasingly proximate goals (successful decision making of some kind) we do not ever reach a legitimate explanation of an organism's behavior in terms of achieving the consistency, coherence, transitivity, etc., that internalist rational criteria demand. Goals may be proximate to varying degrees, but never entirely divorced from the ultimate goal which all natural adaptive behavior subserves. These internalist criteria may, to a certain extent, be characteristic of successful decision-making behavior in a particular environment, but they are not the decision-maker's goal, merely a side-effect of its being well-designed to achieve whatever that goal may be. A decision-maker's deviation from these rationalist tenets will therefore not *necessarily* result in its reduced ability to achieve its goals, since the prescriptions of internalist rationality and the goals of a decision-maker are not coincident.

For instance, in order to meet the criteria of classical rationality, one's preferences must be transitive, that is, if one prefers A over B, and B over C, one must prefer A over C to remain rational. Reinforcement training of various animals demonstrates that they spontaneously develop novel transitive preferences when trained to make pairwise selections between items with adjacent ranks on some arbitrary scale (Delius & Siemann, 1998). That is, when trained to prefer A over B, B over C, C over D and D over E they spontaneously preferred B over D, despite these two items having been reinforced equally over the course of the training. Whilst these data suggest that the mechanism governing the learning of preferences embodies the principle of transitivity, reanalysis of the original reinforcement experiments reveals that simple associative learning rules can account for the ability. The authors conclude that the "capacity for transitive responding could thus be an example for [sic] a trait that has primarily evolved by *exaptation* rather than *adaptation*" (p.131, emphasis added) by which is meant that the selective pressure to discriminate similar stimuli may account for the transitive preferences of pigeons, rats, and humans, rather than any advantage they gain from transitive preferences *per se*. Indeed one can find examples of *intransitivity* in the untrained preferences of animals, as shown in the work of Shafir (1994) on the responses of foraging honey bees to artificial stimuli.

2.2 Optimality and Analysis

Friends of classical rational norms will respond at this point that these norms were never intended as prescriptive rules, but as descriptive tools. Since optimal performance will be achieved by an agent following the prescriptions of classical rationality, they serve a useful purpose in providing the means to calculate a benchmark against which natural performance may be measured. Whilst we as scientists can calculate this benchmark, there is no claim that cognitive mechanisms perform any such calculation. The behavior generated by a rational cognitive mechanism is, however, expected to be well described, or at least approximated, by such optimal models.

We have no objection to this use of optimality modelling. However, it must be pointed out that from this perspective, the discovery that human reasoning fails to meet the internalist criteria of rationality in some situation (whether it be experimental or naturally occurring) should not necessarily be the cause for concern that it has appeared to be within the judgement and decision-making literature (e.g., Kahneman, Slovic, & Tversky, 1982). If internalist rational criteria were never expected to be *implemented* by cognitive mechanisms, but merely to describe their proper behavior (c.f. Anderson's rational analysis), why should one expect arbitrary laboratory test items or natural but novel scenarios to provoke rational responses (c.f. Kahenman and Tversky's heuristics and biases)?

Indeed, a tradition exists within behavioral ecology which treats experimental results not as revelatory of an animal's rationality, but as indicative of its evolutionary history. For example, the field of optimal foraging theory (Stephens & Krebs, 1986) experimentally assesses the foraging behavior of various species in an attempt to discover not whether they are smart or stupid, or rational or irrational, but what the selective pressures on foraging ability must historically have been for these species, and what results these pressures have had in terms of the cognitive adaptations which these species possess. When confronted with what, by the lights of internalist criteria, must be considered irrational behavior, rather than noting the irrationality of the organism involved, these scientists search for environments in which (and adaptive goals for which) sacrificing the missing elements of classical rationality makes sense.

The contrast between the approach of behavioral ecologists and that of decision-making psychologists is crystalized in their response to the possibility of "inappropriate" probability matching in animals and humans (Goodie, Ortmann, Davis, Bullock, & Werner, 1999). The probability matching phenomenon is most straightforwardly presented in a case in which two sites which vary in the rate at which they yield food are attended to in proportion to these yields. Maximizing the consumption of food would be achieved by attending solely to the most productive food source. However it is commonly held that animals and humans often split their attention between the sources in proportion to the expected rate of reward at each source (e.g., Davison & McCarthy, 1988; Tversky & Edwards, 1966). Whilst learning theorists and behavioral ecologists have worked towards discovering in which situations such behavior is successful and adaptive and in which it is not (Williams, 1988), decision-making psychologists have taken the probability matching phenomenon to be evidence of human irrationality (e.g., Dawes, 1988).

2.3 Rational Synthesis

Furthermore, although cognitive mechanisms can be expected to *approximate* optimal solutions to the problems they have been adapted to, we cannot assume that they are also built from approximately rational building blocks. Once we have set aside optimality theories as a means to derive the contents of organisms' heads, it is difficult to see immediately what assumptions are justified when postulating the mechanisms which underpin adaptive behavior. An example from the study of vision highlights this problem.

David Marr (1982) and J. J. Gibson (1979) developed contrasting approaches to solving the problem of how animals achieve visual perception. Marr's computational approach yielded the pipeline model, comprising a series of modules each charged with performing a subpart of the entire task. Each subpart was considered by Marr to be the logical requirement of a system able to form a model of the world around it on the basis of an impoverished two-dimensional array of intensity values (i.e., light falling on a retina). In contrast, Gibson's ecologically inspired theory of direct perception concentrated on how the problem of vision was intimately linked with the problems of acting in the world. For Gibson, the task of vision was not to construct a three-dimensional model of the world from poor quality data, but to reveal the "affordances" of the environment in which the agent was located by exploiting invariants in the rich spatio-temporal visual array.

However, whilst Marr's system was buildable and hence testable, Gibson's theory offered almost no clues as to what might constitute the subparts of visual systems. Alluding to "resonating structures" did nothing to operationalize his theory, which suffered as a result. For our present purposes, what is interesting about this example is that Gibson's ecological considerations did not directly suggest candidate mechanisms in the same way that Marr's computational approach did. Without first principles from which to derive the contents of people's heads, from what source are we to postulate candidate cognitive mechanisms?

In Marr's approach we can glean a clue as to a way forward. Although the processes involved in the pipeline were considered to be the logical precursors to establishing a three-dimensional model of the system's surroundings which could be passed to a suitable spatial reasoning system, Marr did not derive the structure of the pipeline entirely from first principles. Rather, several important empirical results from the neurobiology of vision (e.g., Hubel & Wiesel, 1959) inspired the design of some of the building blocks from which the pipeline was constructed. Once Marr grasped the properties of single cells in the cat striate cortex, for example, he was able to use this understanding to construct edge-detection algorithms. The design process was thus largely data-driven. Indeed the later stages of Marr's pipeline were never adequately modelled due in part to a lack of empirical data with which to inform their design. Like Marr, we must look to empirical studies to suggest candidate cognitive building blocks.

More generally, evolutionary processes can be expected to build complex cognitive as well as perceptual systems from combinations of building blocks, themselves the adaptive result of selective pressures. As such we should conceive of cognitive innards as assemblies of limited cognitive subparts, tinkered with and reassembled by mutation and selection until they fit the environment to which evolution has adapted them. By using empirical evidence from the study of adults, children and other species (e.g., Cummins & Allen, 1998) to suggest the structure of candidate cognitive building blocks, and then exploring how the behavior of various combinations of these building blocks varies with the structure of their environment, we can explore the behavior of model cognitive systems from the bottom up, rather than the top down (Gigerenzer & Todd, 1999).

For example, the recognition heuristic (Goldstein & Gigerenzer, 1999) is predicated on a fundamental psychological phenomenon, recognition memory. This phenomenon has been well studied by psychologists and animal behavior researchers. It clearly subserves much of our everyday behavior. The recognition heuristic utilizes recognition memory to guide decision-making behavior by exploiting the fact that recognition tends not to be randomly distributed across possible entities, but is typically concentrated on the most important ones. The heuristic can be stated as: "A recognized option should be considered to be higher than an unrecognized option on any important dimension". This is clearly a very simple rule. It can be considered to be a building block in that it is informationally self-contained and can act as a subpart of larger cognitive strategies (e.g., Take The Best Gigerenzer & Goldstein, 1999).

This process of *rational synthesis*, the recombination of empirically validated cognitive building blocks, has a counterpart in the field of behavior-based robotics (Brooks, 1991a, 1991b). Increasingly, roboticists interested in building intelligent control systems are coming to realize that problems which appear intractable from the perspective of control theory can be tackled effectively by assembling networks of competing and cooperating behavioral modules. Rather than providing this system with some governing module responsible for coordinating the behavior of these subparts (a fearsome design problem), the robot designers rely on interactions between the robot and its environment to organize the robot's behavior. Although discovering an appropriate combination of modules is not a trivial task, initial successes in both handcrafting (Brooks, 1991a) and artificially evolving (Cliff, Harvey, & Husbands, 1993) such robots suggest that this approach to understanding the design of complex systems is fruitful.

In addition, roboticists interested in using robotic systems to model natural systems have discovered that building robots from empirically validated building blocks can lead to new and interesting theories of animal behavior. Webb (1994, 1996) reports the use of a robot cricket to demonstrate that the phonotaxis achieved by natural female crickets when they approach calling males can be achieved with practically no cognitive mechanism at all, through relying on the acoustic properties of the cricket's ears.

A repeated finding within these related fields is that complex adaptive behavior can arise from the interaction between simple mechanisms and their environment. This observation formed the basis for Valentino Braitenberg's (1984) synthetic epistemology, the use of artificially constructed systems (in this case hypothetical ones) to explore the minimal properties required of systems before various intentional attitudes (fears, desires, beliefs, etc.) are attributable to them. The rational synthesis we employ involves the construction of artificial reasoning systems which are computationally undemanding, and hence psychologically plausible, from decision-making building blocks which are themselves computationally undemanding, and hence psychologically plausible. We then explore the manner in which the performance of such reasoning systems is dependent on facts about the tasks they face and the structure of the environment in which they find themselves.

2.4 The Threat of Exploitation

Our concern with the explanatory role of environment structure in accounting for the performance of a candidate cognitive mechanism has lead us to reject internalist rational criteria as unnecessary for such explanations. For example, what use is transitivity across all choices a cognitive mechanism could ever be expected to make, for instance, if this transitivity is achieved at the expense of adequate response time on a few crucial choices? Might it not be better to sacrifice this property within a set of trivial choices in order to guarantee high speed judgements in a few do-or-die situations?

Yet the employment of internalist rational criteria in the judgement and decision-making literature is commonplace. Why is this the case? One answer is that if the domain to which a cognitive mechanism is expected to apply is unstructured, as it is by definition for the Laplacean Superintelligence, and as is often implied by the use of a flat accuracy performance metric, then environmental considerations will appear superfluous. If success over here is just as good as success over there, then general performance wins out. A corollary of this position is that any failure of reason is equally damaging to a decision-maker's performance. Irrationality will be punished, since disregarding internalist criteria of rationality will leave one open to exploitation. However, limited, structured domains make salient the fact that internalist criteria are obviated when performance on a limited and structured set of items is all that is expected of a cognitive mechanism.

Whilst an organism which fails to adhere to some internalist maxim exposes itself to exploitation in the form of an appropriate money pump or Dutch book (Schick, 1986), for example, if no such exploitatory device exists within the organism's environment, or if the losses due to exploitation are more than made up for by the gains made in other situations, then there is no force to the internalist exhortations. In contrast, if there does exist an exploitatory entity leeching the irrational organism's utility and the organism's irrationality does have net negative consequences on its fitness, then one need not appeal to internalist criteria to demonstrate its irrationality. In this instance, the organism will be irrational by the lights of externalist ecological considerations — it will be unfit.

2.5 Summary

To recapitulate, since organisms are adapted to fit their environment by selective pressures, behaviors and the mechanisms which produce them are only intelligible *in context*. Cognitive mechanisms are bespoke mechanisms, tailored to fit particular circumstances, they are "made to measure". Whilst there may be general trends in dress-making or tailoring (i.e., preference for economy, goodness of fit, quality of material, etc.), these are mere trends, not laws or *a priori* truths. In the same way that opulent, wasteful, ill-fitting, uncomfortable clothing can be fashionable in certain circumstances, so inconsistent, intransitive, seemingly "irrational" cognition will often be adaptive in particular *structured* environments. As researchers we must find ways of appreciating the manner in which a cognitive mechanism's niche is reflected in its structure — we too must be "made to measure" environment structure.

In the remainder of the paper we will explore two important kinds of environment structure which are well-defined and hence measurable. Frequency structure describes the relative prevalence of different decision items within a decision domain. Significance structure describes the relative importance of different decision items within a decision domain. Each class of structure will be explored through manipulating the structure of an artificial decision problem and observing the impact this manipulation makes on the performance and structure of appropriate decision-making heuristics. These rather specific examples are carried out here in sufficient detail to demonstrate the sort of analytical effort that is often necessary to begin to understand why a particular decision mechanism fits a particular environment. They also illustrate some more general lessons about environment/agent interactions and the nature of ecological rationality as a whole.

3 Frequency Structure

We define the frequency structure of a decision-maker's environment as the relative frequency with which each test item is encountered by the decision maker. A flat frequency structure implies that no test item is more likely to be encountered than any other. In contrast, a skewed frequency structure implies that some items are more likely to be encountered than others.

3.1 The German Cities Problem

Here we employ an arbitrary data set (first reported by Gigerenzer, Hoffrage, & Kleinbölting, 1991) as an arena in which to explore the effects of varying frequency structure. The German Cities Problem is an inference task concerning the population sizes of a set of German cities. The task is to judge which is the larger of a pair of German cities. The cities involved are the 83 largest in Germany (all cities with population above 100,000 inhabitants in 1988). The information upon which the judgement must be based consists of nine binary cues (see Fig. 1), for instance, whether the city has a soccer team in the top league of the Bundesliga (the German football league).

This task has previously been used as an inference problem with which to assess the performance of a range of decision-making heuristics (Gigerenzer et al., 1991; Gigerenzer & Goldstein, 1996, 1999; Hertwig, Hoffrage, & Martignon, 1999). However, this previous research has proceeded with no attention to frequency structure, assuming that each comparison between a pair of cities occurs with equal frequency and thus contributes equally to a measure of decision-making performance.

Gigerenzer and Goldstein (1996) report that the recognition rates for these cities (i.e., the proportion of people claiming to recognize each city) increases with population size. On this basis we might assume that the actual frequency structure of this pairwise comparison task (if people encounter this problem at all) is not flat, but that high population cities tend to be reasoned about more frequently than low population cities. This is clearly one manner in which the German Cities Problem environment could be structured. We explore this and a second class of frequency skew, along with their complements, by varying which pairs of German cities are more likely to be encountered:

1a. Product Skew: The likelihood that a pair of cities will be encountered by a decision-maker is proportional to the *product* of the city population sizes.

1b. Reciprocal Skew: This is the complement of Product Skew, the frequency with which a pair of cities will be encountered being *inversely* proportional to the product of the city population sizes.

2a. Similarity Skew: The likelihood that a pair of cities will be encountered by a decision-maker is *inversely* proportional to the difference between the city population sizes.

2b. Difference Skew: This is the complement of Similarity Skew, the frequency with which a pair of cities will be encountered being proportional to the *difference* between the city population sizes.

Whilst we do not know whether one or any of these frequency structures characterizes the distribution of city-size comparisons that people might naturally face, these classes of skew have been chosen because each is probably representative of *some* natural problems. For example, if one encounters entities (cities) at a rate proportional to their value on some dimension (population size), then Product Skew will describe the frequency structure of pairwise comparisons between encountered entities. Similarly, if comparisons between very different entities are handled by some crude early filter, the distribution of remaining comparisons will be biased towards pairs of similar entities. A mechanism operating on this subset will be subjected to a decision environment with a Similarity-Skewed frequency structure. Red deer, for instance, assess the fighting ability of potential opponents by using increasingly sensitive measures (Clutton-Brock & Albon, 1979). The challenger and harem-holder first roar at each other. If there is a significant difference between the volumes, the quieter stag retreats. If roaring fails to decide the contest the stags proceed to the next cue: parallel walking. If this cue also fails to distinguish the stags, they proceed to head-butting. Decision-making mechanisms occurring late in such a sequential assessment will tend to have to distinguish between more similarly matched opponents than those employed earlier in the sequence.

For each class of frequency structure, we explore two degrees of skewness.

1. Mild: The most frequent city pair occurs 10 times more often than the least frequent.

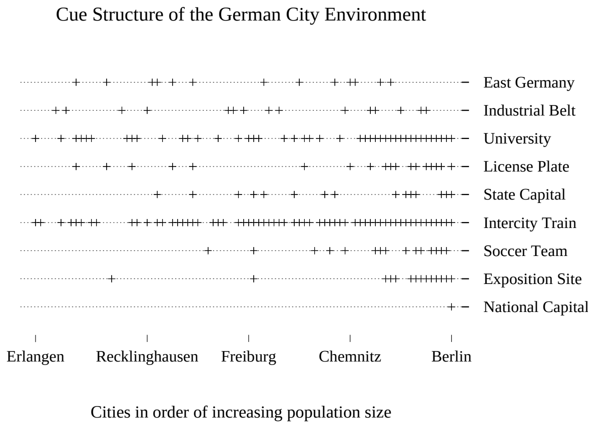

Figure 1: Each city either possesses (+) or does not possess (. . .) each of nine binary cues. Cities in possession of any cue tend to have a larger population size than cities lacking that cue.

2. Extreme: The most frequent city pair occurs 100 times more often than the least frequent.

In each environment the least frequent city pair occurs 10 times. For each environment, the proportion of comparisons in which each individual city takes part is shown in Fig. 2.

It is important to note that, rather than being interested in the problem of comparing city sizes itself, we are concerned with the influence of frequency structure on decision problems in general (but to make our points we will concentrate in depth on this one particular example). Indeed the German Cities Problem is one with perhaps little intrinsic import, serving here as a model, rather than an object of enquiry in its own right.

3.2 The Decision Algorithms

To explore the impact of environmental frequency structure on the structure and performance of decision mechanisms, we chose a small set of such mechanisms for comparison. The four mechanisms we use all make their choices on the basis of some set of the available cues, but they vary in the exact number of cues used and in the complexity of cue processing. The most sophisticated algorithm is multiple linear regression, which first computes the optimal weights for weighting and combining (summing) all of the available cues so that the total difference (error) between the algorithm's predictions (here predicted population size) and the actual criterion values (actual population size) is minimized. Then, to make each individual choice between a pair of objects (e.g. cities), predictions are made for the criterion value of each object by weighting and summing its cue values, and the object (city) with the higher predicted criterion value (population size) in the pair is then chosen as the final decision outcome. Multiple regression thus uses all available information (cues), and is sensitive to their predictive relationship to every object.

The second algorithm, called Dawes's Rule, similarly uses all of the available cues, but it processes them in a rather less sophisticated fashion. Initially, the algorithm must compute the direction of association between each cue and the criterion value — that is, does the cue on average indicate a higher or a lower criterion value (so for example, does having a Bundesliga soccer team indicate a higher city size in over half of the city comparisons?). Then, to make each individual pair comparison, the number of negatively-associated cues for each object is subtracted from the number of its positively-associated cues to create a final score or tally, and the object with the higher score is chosen. This simple method works surprisingly well — Robyn Dawes, after whom it is named, has demonstrated its ability to come close to the performance of multiple regression (Dawes & Corrigan, 1974) — even though it is sensitive only to the "direction" in which each cue points (indicating higher or lower criterion values), but not how strongly.

The last two algorithms take a different approach to decision making. Rather than combining all of the available cues in some manner, they consider cues one at a time, sequentially, until the first cue that enables a decision to be made is found. This decisive cue will be the first which discriminates between the two objects being compared, i.e., one object possesses the cue whilst the other does not. If possession of the cue is positively correlated with the criterion, the object in possession of the cue is chosen. If possession of the cue is negatively correlated with the criterion, the object lacking the cue is chosen. Once a decision has been reached in this way the decision-making process is at an end — all further cues are ignored. Thus all of the available information need not be (and usually is not) considered, let alone processed — and the ultimate decision is always made on the basis of just one discriminating cue. By considering the cues in different orders, different *one-reason decision making* heuristics can be built (see Gigerenzer, Todd, and the ABC Research Group, 1999, for further details).

In particular, here we use the Take The Best algorithm, which orders cues by their validity — that is,

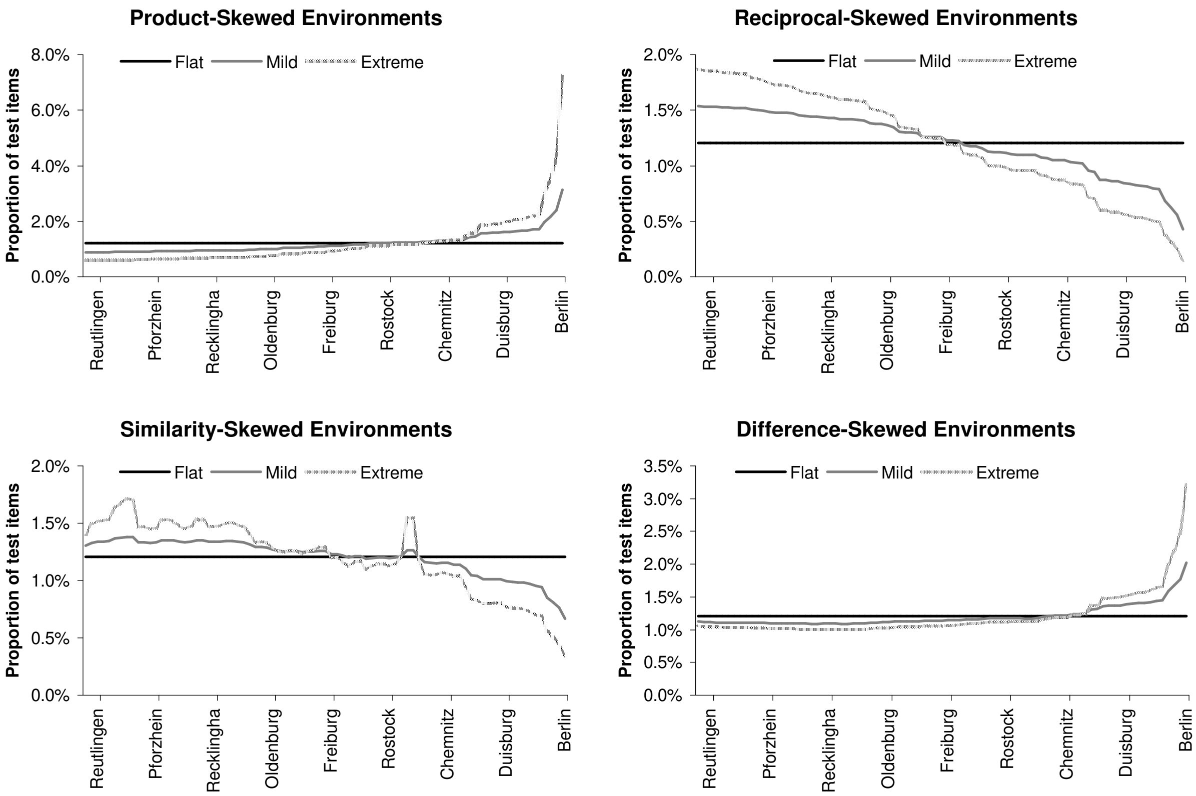

Figure 2: The distribution of test pairs across city size is shown for each of the four frequency-skewed environments in comparison to the default flat environment. The 83 German cities are arranged on the x-axis in order of increasing population size. The proportion of test pairs featuring each city is plotted on the y-axis.

by how often they indicate the larger criterion value in a pair of discriminated objects — so that the best cues are considered first. Take The Best is thus sensitive to the direction *and* strength with which cues indicate the criterion values, but its sensitivity to cue strength only extends to their ranking, not to their precise differences in strength. (Hence, the strength or validity of two cues could change significantly without affecting how they are used by Take The Best; only if their relative ranks change — if one cue becomes stronger than the other — will they be used in a different order.)

We also compare the effectiveness of an even simpler one-reason decision algorithm, the Minimalist heuristic, which examines cues in a random order, stopping when it stumbles upon the first cue which discriminates between the objects. Minimalist is thus not sensitive to cue strength at all, but only to what direction the cue points with respect to the criterion (that is, whether it indicates higher or lower criterion values, or in other words, whether its validity is above or below 0.5). And yet despite its extreme simplicity, Minimalist does not fall far behind the other algorithms, as we will see in the next section.

3.3 Their Performance

Each algorithm was parameterized (e.g., cues ordered or weighted) on the basis of the skewed environment within which their performance was to be assessed. This ensures that each algorithm was appropriately matched to its environment. Each algorithm was then made to judge which was the larger of every possible pair of cities and their average performance across the entire set of pairs was computed. However, some pairs of cities were presented multiple times according to the frequency structure of the environment. Thus, a judgement concerning a frequent pair of cities contributes more to the performance of an algorithm than a judgement concerning an infrequent pair of cities.

Whilst the performance of each algorithm relative to the others remained stable across environments, the absolute performance (i) increased with increasing Product Skew, (ii) increased with increasing Difference Skew, (iii) decreased with increasing Similarity Skew, and (iv) decreased with increasing Reciprocal Skew (Fig. 3). It appears reasonable that choosing the larger of two similarly sized cities will be harder than making the same judgements concerning pairs of dissimilar cities, and perhaps that inferring the larger of a pair of smaller cities will be harder than inferring the most populous of a pair of larger cities, since smaller cities may resemble each other more than larger ones. Because we are dealing with an artificial decision problem we are in a position to move beyond these intuitive assessments of difficulty and explore explanations for the variation in performance caused by our manipulations of the problem's frequency structure.

The source of changes in the algorithms' performances clearly lies in changes in both the predictive validities and the discrimination rates of the cues made available to the algorithms (Fig. 4 and Fig. 5). Validity is defined as the ratio of the number of correct judgements made by a cue to the total number of judgements made by a cue, whilst discrimination rate is defined as the ratio of the number of judgements made by a cue to the total number of judgements sought from a cue. Some cues respond positively to a certain frequency skew, tending to correctly predict a greater proportion of judgements as those comparisons that the cue discriminates correctly become increasingly over-represented. In contrast, other cues may suffer from the same frequency skew, as the comparisons that they deal with correctly become increasingly under-represented. In terms of both changes in validity and discrimination rate, groups of cues appear to respond similarly to particular manipulations of frequency structure, suggesting that a *typology* of cues could be constructed.

In summary, we have seen that frequency structure affects the performance of decision-making algorithms. Despite algorithms having been configured to suit each structured environment, systematic differences in their performance were induced by skewing the frequency structure of these environments in particular ways. The general drop in performance induced by Similarity Skew and Reciprocal Skew coupled with the general increase in performance induced by Product Skew and Difference Skew indicate that the former are harder to deal with than the latter. Whilst different cues respond differently to different frequency structures, the character of this response is often shared by several cues.

3.4 Explaining Performance in Terms of Environment Structure

There are several possible explanations for the changes in performance induced by changes in environment structure in this domain, some of which are specific to the German cities problem and some

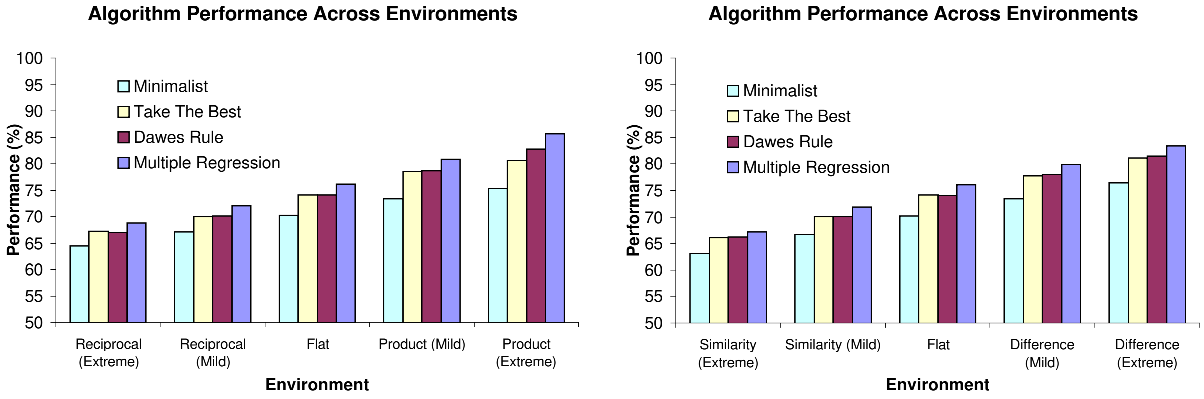

Figure 3: The performance of the four simulated algorithms in each of the 8 structured environments is plotted in comparison to the default flat environment (middle of each panel). Whilst the relative success of algorithms with respect to each other changes little, their overall performance is dependent on the type and degree of skew exhibited by the environment.

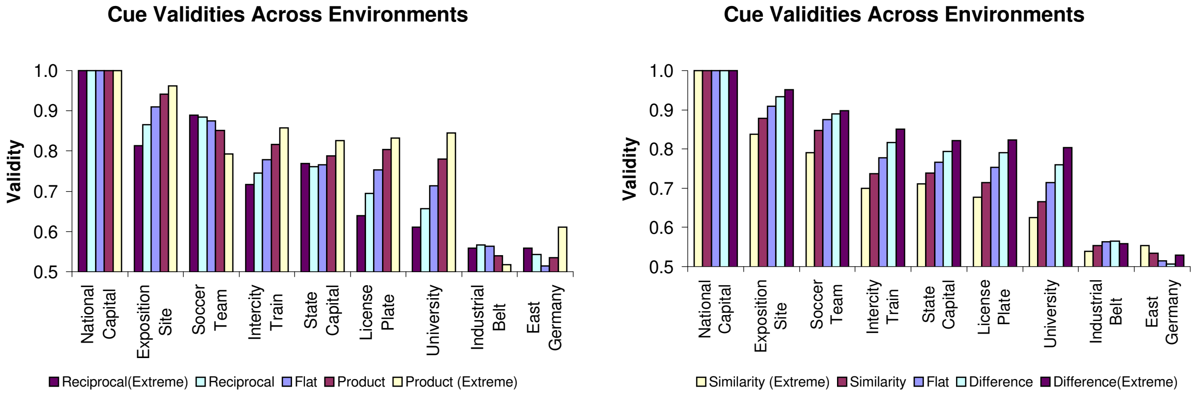

Figure 4: Cue Validity, calculated across all test pairs as (number of correct judgements)/(number of test pairs), varies with environment structure. A cue which correctly predicts a frequent test pair enjoys higher validity. Groups of cues respond similarly to changes in frequency structure.

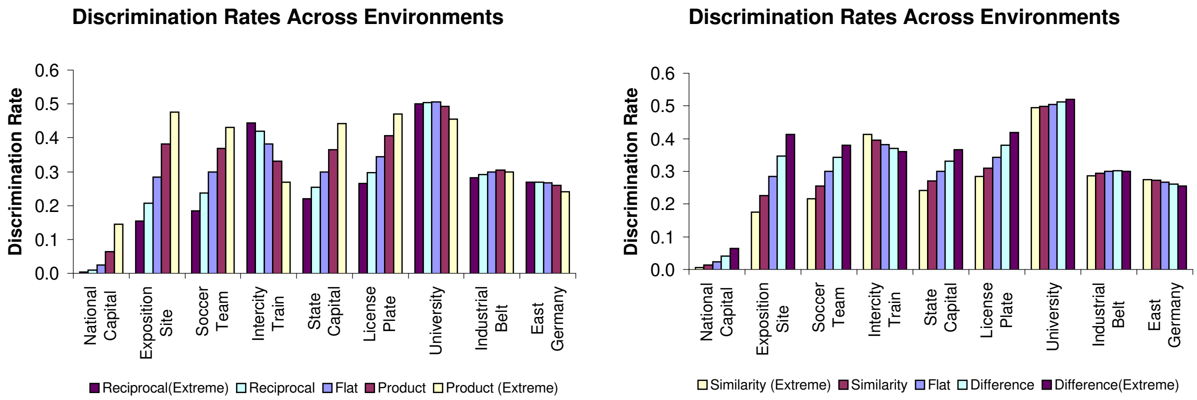

Figure 5: Discrimination Rate, calculated across all test pairs as (number of discriminations made)/(number of test pairs), also varies with environment structure. A cue which discriminates a frequent test item enjoys a greater discrimination rate.

of which are more general. The first and most general is that there exist properties inherent to dichotomous-cue pairwise choice problems which imply that particular kinds of frequency structure will be more difficult than others. The second is that underlying properties of the decision criterion of this particular problem control the impact of different frequency structures. Third, the distributions of the cues across the German cities might influence the manner in which frequency structure affects decision-making performance. Fourth, the changes in environment structure may not make the problem easier or more difficult in general, but either favor or disfavor certain algorithms, of which the ones tested are examples.

A combination of these explanations seems most likely to account for the results reported above. However, it is worth noting some points in favor of this first explanation. Most importantly, all four decision heuristics responded similarly to the changes in environment structure that we imposed. These heuristics differ in many ways, yet benefit or suffer from the same kinds of environment structure. Furthermore, the effects on performance induced by changes in environment structure occur irrespective of the sensitivity of the algorithms to these changes.

For example, the simplest of the strategies tested, Minimalist, is affected by changes in environmental frequency structure, tracking the performance of the other algorithms (although always at a slight distance), despite it not being sensitive to most of these changes. Recall that Minimalist uses the polarity of each cue to govern its inferences. As such, this strategy treats environments identically unless the polarity of at least one cue differs between them (e.g., a cue which predicted high population size in the flat environment predicts low population size in the skewed environment).

For the frequency structures explored here, out of the nine cues in eight skewed environments only five reversals of predictive validity occurred. Four of these reversals affected the East Germany cue, whilst the remaining one affected the Industrial Belt cue. The two cues which suffer validity reversal have the lowest validity of the nine available, ensuring that their reversal makes little impact on the performance of the algorithm. This is not surprising since the polarity of cues with poor validity will be more easily reversed by manipulation of an environment's frequency structure. These observations suggest that Minimalist is typically oblivious to the manipulations of frequency structure that we have imposed on the German Cities Problem.

This type of analysis draws attention to the sensitivity of algorithms to changes in their environment. Minimalist and Dawes's Rule only accommodate changes in the *polarity* of cues. Take The Best is only sensitive to changes in cue validity which are large enough to cause changes in the rank order of cues by their validity (see Fig.6). Multiple regression is in principle sensitive to any change in cue validity. Given these facts, it is understandable that the difference in performance between a sensitive algorithm and a less sensitive algorithm increases with the performance of the former, i.e., the degree to which a sensitive algorithm outstrips its less sensitive competitors increases with the degree to which the sensitive algorithm can exploit the structure of its environment.

This can be seen by looking at the difference between the performance of the most sensitive algorithm, multiple regression, and that of the least sensitive, Minimalist, across all nine environments. This difference is impressively positively correlated with the absolute performance of multiple regression ($r=0.92$). That is, multiple regression benefits from its greater sensitivity to environment structure by exploiting this structure to a greater extent. Indeed all 6 such comparisons between algorithms are correlated in the predicted direction ($r>0.75$) except that Take The Best's advantage over Dawes's Rule in terms of sensitivity does not translate into an increasing advantage over Dawes in the most structured environments ($r=-0.6$).

It is important to stress at this point that we are considering here only the *fitting* performance of the four algorithms — that is, how well they can exploit the structure in a particular set of data from a particular environment. (This situation is also described as one in which the data set on which the algorithm is trained is the same as the data set on which the algorithm's performance is tested.) In this case, the set of data being fitted by the algorithms is the entire frequency-skewed set of all pairs of cities. Thus, there is no *generalization* to new data (where the training set and testing set differ) in the analysis we present here. Generalization performance is of course also of great interest (see Martignon & Schmitt, this issue, for a detailed discussion of the generalization robustness of simple algorithms including Take The Best). But first we must understand more about how the structure in a particular set of data can be exploited by algorithms to make accurate decisions in that same data set.

How can we test whether the changes in performance induced by our manipulation of frequency

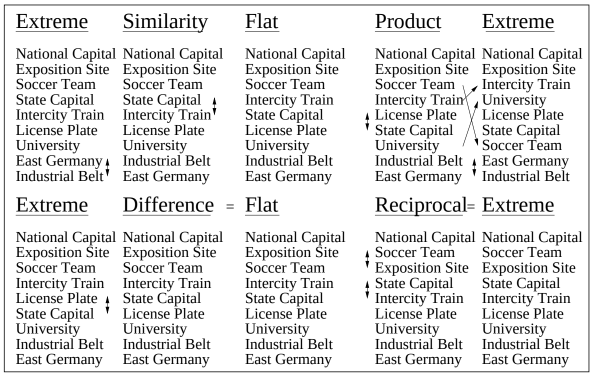

Figure 6: Take The Best utilizes cues in an order determined by their validities. Here the manner in which this cue order changes as the result of frequency structure affecting cue validities is shown for the 9 environments.

structure are not due to some facts peculiar to German cities (and other related environments)? One clue comes from Fig. 2, where it appears that Product and Difference Skew have qualitatively analogous affects on the frequency with which different cities appear in test items. Similarly, Reciprocal and Similarity Skew have comparable effects on these distributions. This could presumably account for the similarity in performance of algorithms in these environments — but how could this pattern arise?

The similarity between Product- and Difference-Skewed environments, and between Reciprocal- and Similarity-Skewed environments, stems from the underlying structure of the distribution of population size across German cities. Since the population of German cities decreases roughly exponentially with rank, forming a so-called J-shaped distribution (see Hertwig et al., 1999), the largest cities (which feature most frequently in the Product-Skewed environments) are also very different from most of the other cities, and hence feature most frequently in the Difference-Skewed environments. Similarly, the many small cities are similar in size to each other and hence are disproportionately represented in both the Reciprocal-Skewed and Similarity-Skewed environments. (If the frequency structure of each environment had been determined using the rank, rather than the real value, of each cities population size, these similarities would be markedly reduced since they rely essentially on the clustering of smaller cities and the isolation of larger cities along the population size dimension.)

Although pairs of the environments do indeed appear alike in their gross characteristics, Product and Difference Skew differ considerably in the extent to which the most common pairs are over-represented in comparison to the least common pairs. Furthermore, Reciprocal and Similarity Skew differ in that the latter features particular mid-sized cities far more frequently than the former. For example, in the extreme Similarity-Skewed environment, Münster and Mönchengladbach, cities which differ in population size by only two thousand inhabitants, feature in 28% more test items (mostly as a pair together) than the average, and in 62% more test items than they appear in within the extreme Reciprocal-Skewed environment (this accounts for the blips to the right of center of the plot of the two Similarity Skew environments shown in Fig. 2). These differences in environment structure are reflected in the fact that some cues respond differently to manipulations which appear superficially similar. For example, the Soccer cue gains validity under Difference and Reciprocal Skew but loses it under Product and Similarity Skew (see Fig. 4). Thus the apparent similarities between environments are perhaps not enough to explain the manner in which algorithm performance varies with frequency structure.

In line with the second explanation for environmental impacts on performance given at the beginning of this section, it could be the case that the arbitrary set of nine cues upon which the algorithms must base their judgements favor certain city pairs over others. Perhaps we have provided no cues which correctly discriminate between small cities, or between similarly sized cities. This type of explanation draws attention to the fact that in structured environments, not just the predictive validity of a cue, but where that validity stems from in the space of possible problem items, is important. In order to assess the relevance of this argument, we need to know whether the nine cues available to the algorithms in this study are representative of the 2^{83} logically possible ways in which a binary cue can apply to 83 objects.

The space required to plot each of these possible cues is prohibitive, but we can expect to approximate the qualitative results by carrying out the same process for a toy problem of 5 objects, and hence $2^5 = 32$ possible cues (Fig. 7). There are $4 + 3 + 2 + 1 = 10$ possible pairwise comparisons between 5 objects (ignoring order). In order to represent the manner in which a cue's performance is distributed across this space of possible comparisons we plot the lower left half of a 5-by-5 matrix containing the outcome of each comparison. Where a cue fails to discriminate between a pair of objects the cell is left blank; correct discriminations are shown in grey; incorrect discriminations are black. Taking the right angle as the origin, cells are indexed by the coordinate (x, y) with object value on the criterion decreasing with increasing x and increasing with increasing y. This ensures that cells near the right-angle of the triangle represent comparisons between objects with dissimilar values on the criterion (e.g., A vs. E), whereas cells near the hypotenuse represent comparisons between objects with similar values on the criterion (e.g., B vs. C). Cells in the upper corner represent comparisons between pairs of objects which both have high values on the criterion (e.g., A vs. B), whereas cells in the lower-right corner represent comparisons between objects which both have low values on the criterion (e.g., D vs. E).

The first thing to note about the distribution of possible cues is that there are far fewer of them than there are possible ways of coloring the cells of one of the triangles used to represent each cue (i.e., 3^{10}). This indicates that the nature of the problem is constraining the kind of cues that are possible. For example it is impossible for one cue to either deal correctly with all possible comparisons or deal incorrectly with all possible comparisons (i.e., no triangle is entirely grey or black). We can see that whilst cues exist which correctly discriminate large cities from small cities (i.e., correctly deal with cells in the right angle of the triangle) and correctly discriminate amongst large cities (i.e., the upper corner), or small cities (the lower-right corner), there are no cues which correctly discriminate amongst many similar cities (i.e., the cells lying along the hypotenuse of the triangle are never entirely grey). These are facts about binary cues in general, and thus will apply to a wide range of environments.

However, it is clear that this reasoning does not straightforwardly apply when continuously valued cues are available to an algorithm which is capable of using them. One continuous cue is sufficient to accurately discriminate between all adjacent objects. Furthermore, a discrete cue with n possible values is capable of distinguishing between all of the adjacent pairs of n objects. A discrete cue with a valency of $\frac{n}{2}$ is able to correctly make half of these pairwise comparisons without incurring error on the remaining pairwise comparisons between adjacent objects. Two such cues would thus be sufficient to achieve perfect performance on the leading diagonal of a problem's triangle diagram.

With this understanding of the space of possible cues in hand, we are in a position to assess the representativeness of the cues made available to the algorithms in the German Cities Problem. The set of cues used in this problem were collected from relevant almanacs containing data on German cities (Ulrich Hoffrage & Ralph Hertwig, pers. comm., 1999). As such the cues are a relatively representative sample of the kind of facts people might know about cities. The manner in which correct and incorrect judgements are distributed over the space of possible comparisons for each cue is plotted in Fig. 8 according to the same principles described above for the 5-object case. The cues involved in the German Cities Problem tend to allow discrimination amongst the larger cities, and between larger and smaller cities, but fail to discriminate correctly amongst similarly sized cities, or amongst small cities. The first of these deficiencies stems from the logical constraints of pairwise choice and binary cues. As just argued for the 5-object case, there simply do not exist cues which correctly deal with many comparisons between objects with similar values on the criterion dimension. In order to accurately deal with each comparison along the hypotenuse of the triangle diagrams presented here, 83 binary cues must be consulted.

In contrast, the fact that the cues available to the algorithms facing the German Cities Problem do not tend to discriminate amongst small cities, is not a result of some constraint on binary cues. This deficiency is due to this set of nine cues being a biased sample of logically possible cues. Is there an explanation for this bias, or must it be attributed to the vagaries of sampling error? There are reasons to believe that the former is most likely.

Whilst there may exist cues which discriminate amongst smaller cities, they are unlikely to be recorded in almanacs, which, since larger cities are more interesting to their readers, tend to record facts which are true of large cities, and false of small ones. In addition, these facts are not true of every large city, but tend to be false of almost every small city, ensuring that they tend to discriminate amongst large cities as well as between large cities and small cities, but not to discriminate well amongst small cities.

Thus, randomly sampling cues from those made available in the public domain will tend to result in a set of cues which is not representative of the space of possible cues, but which is biased towards those cues suitable to the structure of the problem which they have been selected for. This set of cues will not be able to accommodate a manipulation of environment structure, if this manipulation opposes the natural structure of the problem responsible for their existence in the public domain. In skewing the German Cities Problem in the direction of small population size, we have opposed the natural tendency for large cities to be more frequently reasoned about and discussed. As a result, the validity of cues taken from almanacs has tended to fall under Reciprocal Skew.

This argument does not apply solely to the German Cities Problem, but in principle can be generalized to any decision problem. Well adapted decision makers will tend to recognize and attend to cues which are well-suited to the predictive demands of the problem as influenced by its frequency structure and significance structure. This implies that, to the extent that such cues are logically possible, the cues used by such decision makers will tend to discriminate correctly between frequent and/or significant pairs of objects, possibly at the expense of rare and/or insignificant pairs. However, such a selection of well-adapted cues will not necessarily support performance on a differently structured decision problem. More specifically, if a decision problem is artificially skewed in favour of precisely those items which are insignificant in the natural decision-making problem, natural cues will tend to be unable to cope with this manipulation. For the German Cities Problem, this inability to cope with unnatural problem structure is manifested in the poor performance of algorithms in the Reciprocal Skew conditions.

In concert, the effects outlined above ensure that the structure of the 9 cues made available to algorithms in the German Cities Problem favours environments where they are more often called on to choose between pairs of large cities, or between large and small cities (Product and Difference Skew, respectively). For the same reasons, these algorithms will tend to perform poorly when forced to choose more often between small or similarly sized cities (Reciprocal and Similarity Skew, respectively). These general trends should apply to any binary-cue-based choice environment where alternatives at one end of the criterion dimension are more important or frequent than those at the other end.

In summary, the variation of algorithm performance with environment structure can be traced to several sources. First, some classes of frequency skew are inherently difficult to accommodate due to the nature of binary cues and the pairwise choice paradigm. This argument accounts for the reduced performance on Similarity-Skewed environments. Second, some classes of frequency skew are difficult contingent on the cues available. This argument accounts for the reduced performance on Reciprocal-Skewed environments. Third, some algorithms are more sensitive to environment structure than others and are thus more likely to accommodate particular manipulations. The heuristics assessed here vary in their sensitivity to environment structure, and this sensitivity manifested itself in differences in the size of the advantage one algorithm achieved over another in different environments.

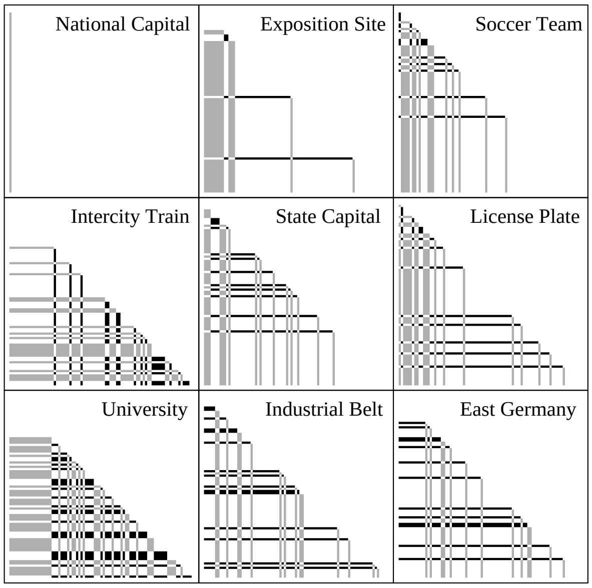

Figure 8: Judgement distributions for each of the 9 cues made available to the decision-making heuristics in the German Cities Problem. Each triangle represents all possible pairs of cities (because pair order is irrelevant, the upper half of each matrix is redundant, and hence omitted). Cities are arranged in order of increasing population size from left to right and top to bottom. Cues are arranged in order of increasing validity in a flat environment. Grey indicates correct inferences, black indicates incorrect inferences, and cues fail to discriminate in the remaining instances.

3.5 Concluding Thoughts on Frequency Structure

By employing the German Cities Problem as a toy environment, we have shown that frequency structure impacts on the performance of decision-making algorithms. The character of this impact is complex. The presence of environment structure

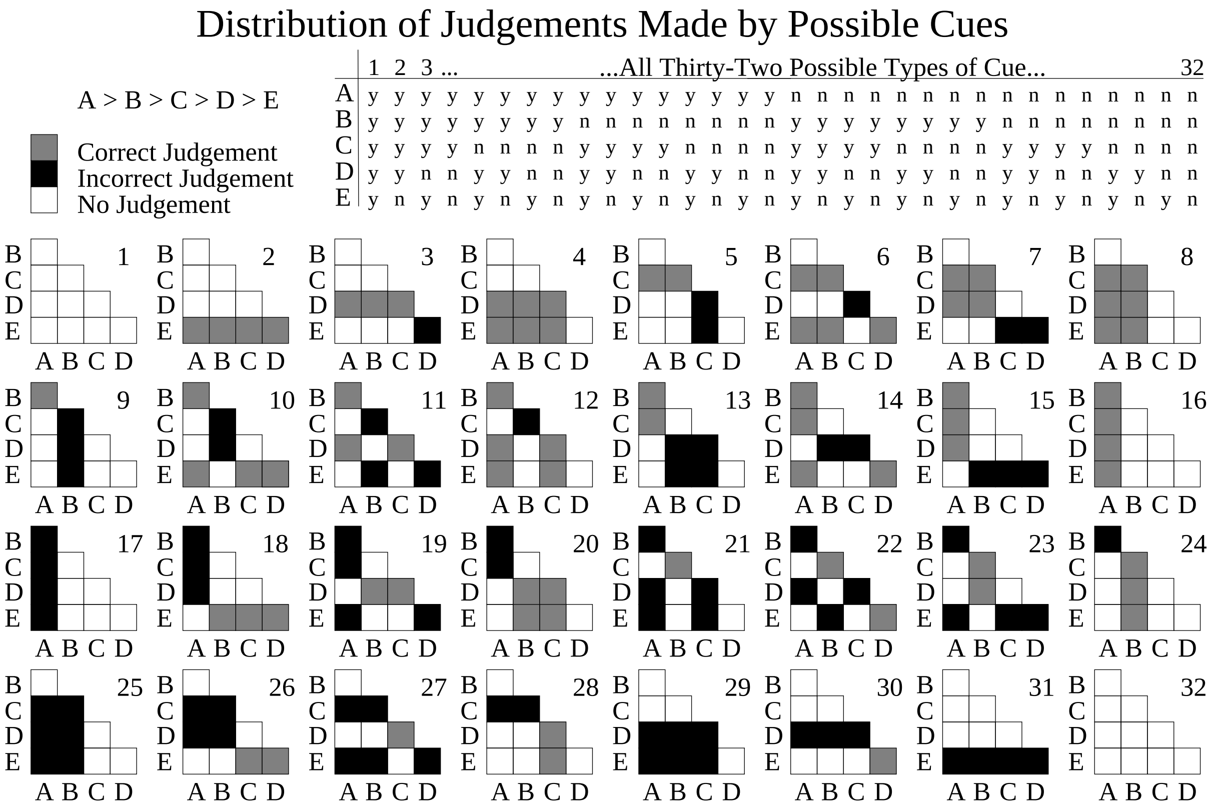

Figure 7: There are 32 possible ways in which a binary cue can apply to five objects, A–E. The order of the objects on the criterion dimension is described by the inequality A>B>C>D>E. Each of the 32 numbered triangles depicts the judgements made between all possible pairs of objects on the basis of the corresponding cue shown in the table above. Judgements may be correct (grey) or incorrect (black); a blank cell indicates that a cue does not discriminate between the pair of cities involved in the judgement.

demands that decision-makers trade off general performance against performance on important subsets of test items. As a result, not only the validity of a cue, but the source of this validity is of importance to decision makers. Cues which gain their validity from frequent test-items are more useful than equivalent cues which gain their validity from rare test-items.

Furthermore, environment structure interacts with the necessary and contingent characteristics of a decision problem, and the strengths and weaknesses of a particular algorithm, to influence the performance of that algorithm.

4 Significance Structure

As well as differing in their relative frequency, naturally occurring problems differ in their relative significance. Consider a list of decisions which might be faced on the way to work: Which tie should I wear? Should I walk to the bus stop or ride my bicycle? Which bus should I catch? Is it safe to cross the road? How fast should I walk? How fast is that car approaching? Should I jump left or right? How am I going to make that 9:15 meeting now?

Clearly these dilemmas differ along many dimensions; some are leisurely, some pressing; some are conscious, some unconscious; some are casual, some weighty. Here we will consider the effects of variation in the importance or gravity of decision problems on the structure and performance of decision-making mechanisms.

There are two ways in which the significance structure of a decision problem can be mischaracterized. First, the goal of the decision-maker may be misconceived. For example, doctors may be assessed on the accuracy of their diagnoses when what is significant to them is not forming an accurate judgement of what ails a patient, but prescribing measures which will alleviate this ailment. Whilst correct diagnoses are clearly a step towards this goal, they do not constitute it. There may be diagnostic errors which have no effect on a doctor's prescription because the confounded conditions demand the same treatment (see Connolly, this issue, for discussion along these lines). Similarly, the prescription of an incorrect treatment regime may, nonetheless, sometimes result in a cured patient (e.g., prescribing a course of vitamin supplements, complete rest and avoidance of dairy products, when the correct treatment was merely relaxation). Mischaracterising the aims of the decision maker leads to a misunderstanding of what counts as success and what counts as error.

The second, and related, manner in which significance structure may be misconstrued is in failing to appreciate that different decision problems differ in their significance to the decision maker, i.e., failing to discriminate between inconsequential decisions and those of much greater significance. A doctor confronted with what appears to be a case of influenza faces a decision problem which differs from that of a colleague encountering what appears to be a case of meningitis. Errors in treating such cases would have radically different consequences. Assuming that a doctor will not treat his patients with 100% accuracy, it is of the utmost importance that the errors which are made are distributed amongst the less important cases rather than those involving life-threatening illnesses. Indeed, it may be necessary to trade off accuracy in general against accuracy over an important subset of decision items (Sober, 1994). Assessing a doctor's performance using a metric which is insensitive to differences in significance will fail to capture this trade-off.

In general, a problem's significance structure is the manner in which the different decision items which constitute the problem differ in terms of their consequences for the decision maker's goal. For dichotomous decision problems such as the ones considered here, in which a test item's significance can be operationalized as the difference in value between the two possible outcomes of the decision regarding that item, significance structure describes the manner in which this difference varies across the space of possible test items.

4.1 The Mushroom Problem

Imagine a fungivorous forager which, throughout its lifetime, encounters mushrooms, one after the other. Whilst some of these mushrooms are good sources of valuable nutrition, others contain damaging toxins. When confronted by a mushroom, the forager must decide whether to eat it, or reject it in favor of a safe but mediocre food source assumed to be ever present in the forager's environment. The forager must make its decisions on the basis of binary cues which it is sensitive to, and which together describe each mushroom, for instance, odorous versus odorless, colorful versus dull, and so on.

The significance of these decisions will vary across the space of mushrooms liable to be encountered by a forager. How will this variation impact on the success of the different foraging strategies that such a forager might employ? In order to answer this question we simulated such a forager, and

Cue	Validity	Hits	Misses	False Alarms	Rejections
odor	0.886	0.419	0.015	0.098	0.467
gill-size	0.757	0.483	0.208	0.035	0.274
bruises?	0.744	0.339	0.077	0.179	0.405
population	0.670	0.240	0.052	0.278	0.430
gill-color	0.666	0.305	0.121	0.213	0.361
spore-print-color	0.661	0.429	0.250	0.089	0.232
habitat	0.624	0.466	0.324	0.052	0.158
gill-spacing	0.616	0.148	0.014	0.370	0.468
stalk-color-above-ring	0.556	0.445	0.371	0.073	0.111
cap-shape	0.555	0.290	0.217	0.228	0.265
stalk-shape	0.553	0.319	0.248	0.199	0.234
stalk-color-below-ring	0.552	0.439	0.369	0.079	0.113
cap-surface	0.551	0.377	0.308	0.141	0.174
stalk-surface-below-ring	0.537	0.082	0.027	0.436	0.455
cap-color	0.533	0.292	0.241	0.226	0.241
ring-number	0.522	0.518	0.478	0.000	0.004
ring-type	0.522	0.518	0.478	0.000	0.004
stalk-surface-above-ring	0.515	0.052	0.019	0.466	0.463
veil-color	0.505	0.024	0.001	0.494	0.481
gill-attachment	0.504	0.024	0.002	0.494	0.480

Figure 9: The appearance of each mushroom is characterized by twenty dichotomous cues. The rates of Hits, Misses, False Alarms and correct Rejections have been calculated across the entire set of 8124 mushrooms. Hits are cases in which a cue correctly indicates that a mushroom is edible. Misses are cases in which a cue incorrectly indicates that a poisonous mushroom is edible. False Alarms are cases in which a cue falsely indicates that an edible mushroom is poisonous. Correct Rejections are cases in which a cue correctly indicates that a mushroom is poisonous. Cues are shown ordered by their Validity, where Validity=Hits+correct Rejections.

Orthodox	Edible	Poisonous
Accept	+1	0
Reject	0	+1

Flat	Edible	Poisonous
Accept	+5	-5
Reject	+.18	+.18

Lethal	Edible	Poisonous
Accept	+5	-∞
Reject	+.18	+.18

Odor	none	almond	anise	musty	creosote	spicy	pungent	fishy	foul	
Accept	+5	+15	+20	-	-	-	-	-	-	Edible
	-5	-	-	-5	-5	-5	-5	-5	-10	Poisonous
Reject	+.18									Edible
										Poisonous

Figure 10: Four payoff matrices determining the significance structure of the Mushroom Problem. Each cell contains the points awarded for an individual decision. Dashes in the Odor matrix indicate that no mushrooms were present in a particular cell.

explored how the performance of various foraging strategies was affected by manipulation of the significance structure of the artificial mushroom environment it inhabited.

We utilized Schlimmer's (1987) database of 8124 different mushrooms from 23 species within the Agaricus and Lepiota families (available from the University of California, Irvine Machine Learning Repository; Blake, Keogh, & Merz, 1998). Each mushroom was described using 20 binary cues (dichotomized versions of the original data), as shown in Fig. 9. Of the 8124 mushrooms, 4208 (51.8%) were classified as edible, whereas 3916 (48.2%) were classified as poisonous. The rates at which each cue is able to distinguish poisonous from edible mushrooms can be captured by four values: Hit rate, Miss rate, False Alarm rate, and Correct Rejection rate. These rates correspond to the cue's tendency to correctly or incorrectly indicate edible mushrooms, and incorrectly or correctly indicate poisonous mushrooms respectively, and are reported in Fig. 9. A cue's validity can be calculated as the proportion of correct inferences it makes, i.e., as the sum of its hit rate and correct rejection rate.

The significance structure of this decision problem can be manipulated by defining different payoff matrices governing a decision maker's performance. Fig. 10 depicts the four significance structures we explored. The first represents a scheme which assumes no significance structure exists. A decision maker receives a point for each positive response to an edible mushroom and each negative response to a poisonous mushroom, and no points for any other responses. This scheme rewards accurate classification and is termed Orthodox since accuracy metrics of this type dominate much of decision-making psychology. A student being tested on his knowledge of mushrooms might be assessed in this way — the student is sent out into the environment with two baskets, one labeled *edible*, one labeled *poisonous*. Upon his return, a teacher awards a point for every mushroom that the student has placed in the correct basket.

This Orthodox significance structure treats all successes as equivalent and commensurate, and all errors likewise. However, a forager actually consuming or rejecting mushrooms has not achieved its goals to the same extent by rejecting a poisonous mushroom as by consuming an edible one. Although these are both appropriate behaviors, in the latter case the forager has gained valuable nutrition, in the former it has avoided being poisoned. Similarly, for such a forager, the consequences of the two classes of possible error differ radically. Whilst the rejection of an edible mushroom incurs an opportunity cost, the consumption of a poisonous one incurs the debilitating effects of whatever toxin the mushroom contains.

The second payoff scheme attempts to capture this significance structure to a greater extent through awarding points for eating edible mushrooms, deducting points for eating poisonous mushrooms, and awarding a negligible amount for rejecting mushrooms in favor of the alternative mediocre foodstuff. The payoff matrix is constructed such that eating all mushrooms achieves, on average, the same score as rejecting all mushrooms. This scheme can be considered to offer the forager the choice between a risky, but potentially high value food item (the mushroom) and a safe, but relatively low value food item (the alternative). It is termed Flat, since each poisonous mushroom and each edible mushroom are equivalently poisonous or nutritious.

The two environments described so far can be adequately captured by a signal detection paradigm. In varying the points awarded for eating and rejecting mushrooms which are poisonous or edible we have been defining the costs and benefits of the four cells in a signal detection matrix — hits, misses, false alarms and correct rejections.

However, significance structure can be finer grained than the signal detection picture implies. In the third environment, termed Odor, the value of consuming edible mushrooms and the cost of eating poisonous mushrooms is correlated with their odor. Whilst the fungivore can discriminate between odorous and odorless mushrooms, the significance of a decision involving a particular mushroom depends on whether the mushroom smells “foul”, “fishy”, “pungent”, and so forth, that is, on features which are not directly available to the forager, but may be recoverable from combinations of the dichotomous cues which *are* available. Within this environment, the costs and benefits of hits and misses vary systematically across the space of decision items.

Furthermore, significance structure can sometimes be difficult to capture in the terms of signal detection. For example, in reality poisonous mushrooms may be more dangerous than the deduction of points implies. The fourth environment is identical to the Flat environment save that the consumption of any poisonous mushroom results in the death of the fungivore, that is, an immediate and irreversible assignment of a score of zero points to the forager. This Lethal environment ensures that successes and failures cease to be measured in commensurate ways. No amount of edible mushrooms can be eaten to offset the consumption of a lethally poisonous mushroom. This is indicated in Fig. 10 by assigning a utility of negative infinity to the miss cell of the Lethal payoff matrix.

Cue	Rule 1	2	3	4	5	6	7
Present	✗	?	✗	?	✓	?	✓
Absent	?	✓	✓	?	✗	✗	?

Figure 11: Each cue is treated in one of seven ways. The presence or absence of a cue can prompt a forager to reject (cross) or accept (tick) a mushroom, or to check the next cue (?). Notice that since, across the entire population of mushrooms, the presence of each cue tends to indicate edibility, high-performance foragers might be expected to utilize rules 5, 6, and 7 more than 1, 2, and 3. Rules 3 and 5 always stop search since they propose a definite action based on the presence or absence of the cue they apply to. Rule 4 ignores the cue it is applied to.

These four environments demonstrate the range of possibilities that a problem's significance structure can cover. Real decision problems can be expected to exhibit significance structures which are more complex still than those explored here since neither options, nor the evidence upon which to decide between them, need be binary in nature; further, differing outcomes may not be as easily reducible to a single dimensional of utility. In the next section we assess the effects that the four variations of significance structure have on the performance and structure of a class of simple decision making algorithms.

We can make some general predictions regarding the effects of these manipulations. For instance, those algorithms tailored to an inappropriate significance structure should tend to be outperformed by those which are appropriately tailored. In addition, algorithms tailored to the Lethal environment should be conservative in their food choice, whilst those tailored to the Flat or Odor environments should tend to make errors within a subset of insignificant mushrooms in comparison to those algorithms tailored to the Orthodox environment, for whom one error is equivalent to any other.

4.2 The Algorithms

Here we explore a class of lexicographic decision algorithms. Like Take The Best, described above, these decision heuristics treat evidence one piece at a time and make a decision based on the first piece of evidence to suggest a course of action other than checking for more information, i.e., the first piece of information that allows a choice to be made. In this case the evidence is in the form of binary cues which are consulted in some order (tied ranks are possible in which case the tied cues are consulted in random order). Each cue is associated with a stopping rule. This rule determines whether the presence or absence of the cue leads to the forager eating or rejecting the mushroom, or to the forager consulting the next cue. We model seven different stopping rules (Fig. 11). If an algorithm checks all 20 cues without making a decision, the action taken is determined by a biased coin toss.

To understand how significance structure can interact with the structure of decision mechanisms and affect their performance, we will focus on this example task to find and compare strategies which perform well within each of the four environments described above. We cannot assess each member of the class of lexicographic rules since, given that cue ranks may be tied, there are over 20! orderings of cues and each ordering can be governed by 7^{20} combinations of stopping rules. To find lexicographic algorithms which suit the Mushroom Problem under a particular significance structure, we implemented a form of parallel search inspired by natural evolution.

The genetic algorithm we used (Holland, 1975; Goldberg, 1989; Mitchell, 1996) started with a population of 1000 randomly generated algorithms and assessed the performance of each on the Mushroom Problem under a particular significance structure (i.e., in a particular environment). Each assessment involved the particular algorithm encountering 100 mushrooms drawn at random from the population of 8124, eating or rejecting each mushroom, and gaining or losing points as a result. Once each of the 1000 algorithms was assessed, a new population of 1000 algorithms was generated by allowing the better performing algorithms to "reproduce", that is, to be copied into the next generation. This copying procedure was subject to a small chance of error which introduced "mutations" into the strategies. The newly generated population of offspring algorithms was then assessed as before and the process was repeated until 5000 generations of simulated evolution have taken place.

As a result of this assessment, reproduction, and mutation cycle, the population of 1000 algorithms became better and better adapted to the problem it faced. Over many thousands of generations performance increased as the algorithms converged on successful orderings of cues and appropriate stopping rules for these cues.

In each of the four environments depicted in Fig. 10 we assessed 20 independent populations of 1000 algorithms each for 5000 generations of simulated evolution. During reproduction, there was a 1 in 100 chance that each of an algorithm's parameters might be mutated. Mutations, when they did occur, consisted of (i) a cue's rank being replaced by a random value drawn from the set $\{0.5, 1, 1.5...20, 20.5\}$[1], (ii) a cue's stopping rule being replaced by one drawn at random from the seven possible rules, or (iii) a strategy's biased coin being replaced by a coin with bias drawn randomly from the range [0,1].

For each of the four environments, the top 5 (0.5%) foragers from each of the 20 populations at generation 5000 were collected, and their long-term mean performance over 10,000 lifetimes (i.e., 1,000,000 mushrooms) was calculated. The best such long-term mean performance was recorded. Algorithms which failed to achieve a long-term mean performance within 5% of this threshold were discarded.

Duplicate equivalent strategies were then excluded. Strategies were deemed equivalent if they exhibited the same cue ordering and applied the same stopping rules to these cues, once redundant cues had been removed. Redundant cues were either those associated with stopping rule 4, those which were never consulted because a cue associated with rule 3 or 5 preceded them in the cue order, or those which, over the course of 10,000 lifetimes, although consulted, had never stopped search. The remaining "elite" strategies are thus unique and perform well in the environment to which they were adapted.

4.3 The Elite Strategies

At this point, we will delve into a specific detailed analysis of the evolved strategies in these environments to see what general principles we can uncover and to demonstrate the sorts of analytic approaches that can aid in such a search. A first indication that the strategies fit for one environment tend to differ from those fit for another is given by the Venn diagram in Fig. 12 which demonstrates that of the 93 elite strategies found through evolutionary search, only 2 occurred in more than one environment.

[1]Half ranks were employed so that cues could mutate to fall in between two previously adjacently ranked cues. After reproduction, ranks were renormalized so that they were again consecutive integers.

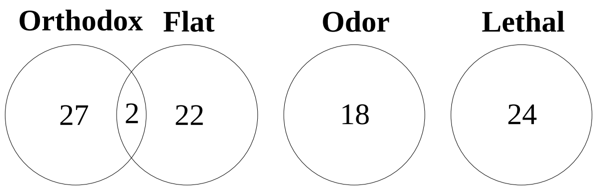

Figure 12: Two elite strategies arose in both the Orthodox and Flat environments. The remaining strategies are unique to the environment in which they evolved.

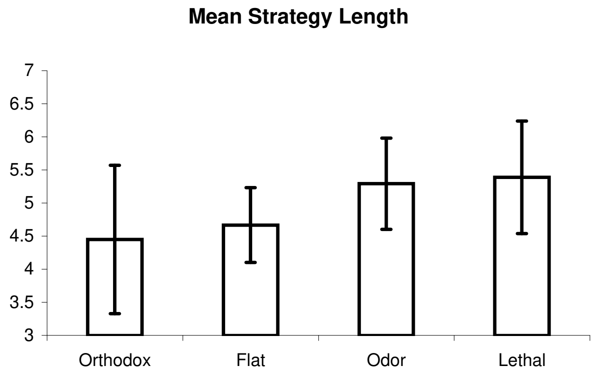

Figure 14: The mean number of cues involved in the elite strategies from each of the four environments explored. This measure differs significantly across the four conditions (χ^2 test, $p < 0.001$). Whilst neither Orthodox and Flat, nor Odor and Lethal differ from one another (χ^2 test, $p < 0.5$), together Orthodox collapsed with Flat differ significantly from Odor collapsed with Lethal (χ^2 test, $p < 0.0001$).

How do the elite strategies within one environment resemble each other, and how do they differ from those found in different reward regimes? The set of elite Orthodox strategies is heterogeneous in that many cues feature across the strategy set, and there is little consensus regarding which cues are useful and which are not (Fig. 13). In contrast, the other three sets of elite strategies each feature a smaller number of cues, and exhibit a higher degree of consensus regarding which cues are important. Furthermore, each *individual* elite strategy in the Odor and Lethal environments tends to involve a slightly but significantly greater number of cues than elite strategies found for the other two environments (see Fig. 14).

In combination, these results suggest that as the significance structure of a decision environment becomes increasingly heterogeneous, i.e., the difference in significance between decision items increases, appropriate strategies become increasingly homogeneous and less frugal in cue use. While the set of elite strategies for the Orthodox environment is wide and shallow, those of the Lethal and Odor environments are narrow and deep. This phenomena is reminiscent of findings concerning the differences between novice and expert decision makers. While novices tend to pursue a variety of strategies and as a group may attend to many different sources of potentially relevant information, experts are less variable in their approach to a problem, typically using just those few specific cues which are most appropriate to the decision problem at hand (Shanteau, 1992).

The particular cues which feature in elite strategies for the Mushroom Problem can be regarded as falling into three groups. First, a few high validity cues (e.g., odor and bruises) show up in nearly every elite strategy, regardless of which environment the strategy has adapted to. Second, a set of auxiliary cues (e.g., stalk shape and gill-spacing) tend to feature in many of the elite strategies within a particular environment, but do not feature strongly in alternative environments. Third, the remaining utilized cues tend to be idiosyncratic to particular strategies within particular environments. It is clear that attending to high validity cues will be a useful part of most any decision strategy, and this observation can account for those cues that are utilized frequently across all environments. However, cues are not always utilized in proportion to their validity, even within the Orthodox environment. Reasonably accurate cues may be utilized only vary rarely. For example, gill-color, which is ranked fifth in terms of validity, is never involved in any elite strategy in any environments.

Similarly, what marks particular cues as appropriate to particular environments can be hard to trace. The spore-print-color, habitat, and stalk-surface-below-ring cues are present in many of the elite strategies evolved within the Lethal environment. However, these cues share few features which can explain their utility. They are mid-ranking in terms of validity. Although the stalk-surface-blow-ring cue enjoys a low Miss rate, which given the significance structure of the the Lethal environment would appear to be crucial to the utility of cues, the other two are unremarkable in this respect. However, spore-print-color and habitat do enjoy low rates of False Alarms. How are we to explain this curious choice of cues?

Given that no cue perfectly predicts edibility across the entire set of mushrooms, no cue can initially be used by a lexicographic strategy to identify

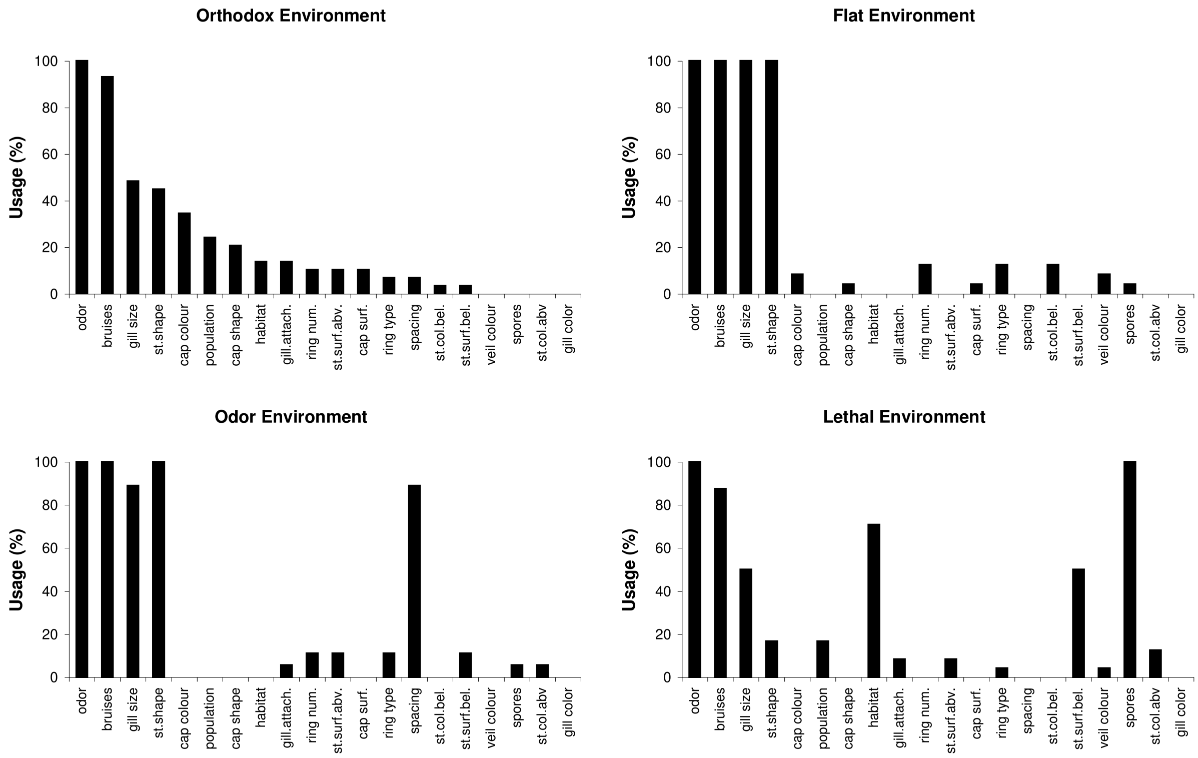

Figure 13: The percentage of elite strategies which involve a particular cue in each of the four environments tested. Cues are ordered as in Fig. 9, i.e., in order of their validity in the Orthodox environment. Notice that a core group of high validity cues are attended to by many elite strategies regardless of environment (tall bars at left), while a number of cues are attended to by many strategies within specific environments (tall bars toward right), and remaining cues may be attended to by individual strategies within an environment (short bars).

edible mushrooms without error. Since the consumption of a poisonous mushroom is fatal in the Lethal environment, every successful strategy there must proceed by rejecting subsets of mushrooms on the basis of cues which tend to make correct rejections and few false alarms. It is in this respect that spore-print-color and habitat (and odor and gill-size) excel, allowing a strategy to confidently reject mushrooms in the knowledge that those unrejected will for the most part be edible. A successful strategy will use early cues of this kind to filter out poisonous mushrooms such that those remaining can be split into definitely edible or possibly poisonous by a subsequent cue (e.g., stalk-surface-below-ring).

However, this rather involved explanation cannot enable us to state in advance which particular cues will be employed within elite Lethal strategies, but merely to offer a post-hoc analysis of successful strategies. Even in this respect the explanatory strategy is imperfect since it cannot account for why alternative cues were not utilized in place of those that were. For example, there exist cues with lower false alarm rates than spore-print-color and habitat which were not employed to any great degree. Why were these cues eschewed?

In the Odor and Flat environments, the distribution of cue usage is even harder to understand. Gill-spacing, a popular cue in the Odor environment, is unremarkable save that it enjoys a low miss rate. However, there is little indication that misses are more crucial in the Odor environment than in the Flat environment, for instance, where the gill-spacing is never utilized by an elite strategy.

The reason for the difficulty we experience in predicting and explaining the successful cue orderings stems from the properties of lexicographic strategies and our reliance on measures of cue performance derived from their application to the entire space of decisions. A strategy's highest ranked cue will be consulted in all decisions. However, since this first cue may sometimes suggest a course of action (i.e., eating or rejecting) other than checking the value of the next cue, this next cue will only figure in a subset of the decision made by a strategy. Similarly, the third cue will be consulted for a subset of this subset — a subsubset of encountered mushrooms — and so on. As a result, characteristics of a cue which have been calculated across the whole environment, even if they suitably accommodate significance and frequency structure, will tend to become less and less useful the deeper into a lexicographic strategy the cue is placed.

Fig. 15 demonstrates this problem by depicting the direction in which cues at each rank in a lexicographic strategy tend to be utilized. Recall that depending on the stopping rule employed in conjunction with a cue, its presence or absence can be the prompt for either positive (eat), negative (reject) or neutral (check next cue) behavior. Rules can be divided into those which tend to consider the presence of a cue to be an indicator of edibility and/or its absence to be an indicator of toxicity, and those for which the presence or absence of the cue indicates the opposite. One might expect that since, on average across the Mushroom Problem data set, the presence of each cue tends to indicate edibility, rules of the former kind might be more useful and hence better represented in the set of elite strategies. Fig. 15 shows that this is indeed the case *early* in a strategy. The first cue used by an elite strategy is always consulted in conjunction with a rule of this expected polarity. However, as we descend through the ranks, more and more of the cues begin to be associated with rules which operate in the opposite direction, until the polarity of a cue across the whole population of mushrooms ceases to be a predictor of rule use at all.

The divergence between the performance of a cue over an entire space of problem items (global validity) and its performance across the subset of items which it actually encounters as a consequence of the cues preceding it in a lexicographic ordering (conditional validity) can be expected to increase with the rank of a cue, as mentioned above. In addition, the rate at which this divergence increases with rank can be expected to itself increase with the degree to which the significance or frequency structure of an environment tends to focus performance on fewer decision items. Consider that in the Orthodox environment, the contribution of each individual success or error on the part of a cue to its validity is equal. In contrast, within the Odor environment, there is a differential contribution of successes and errors to global validity. If a particular mushroom is highly nutritious, then successful cues will tend to be able to identify it as edible. The *global* validities of each of these cues will be inflated by their ability to correctly identify this mushroom. However, the conditional validity of only one cue will be increased by this ability. This is due to the fact that, in practice, only one cue will ever be used to identify this mushroom. The remaining cues which could also have made this correct identification have missed out. As a result, their conditional validities will not reflect their global validities, since whilst the latter measure takes their performance on every decision item into account, the former does not.

This issue closely parallels the problem of model

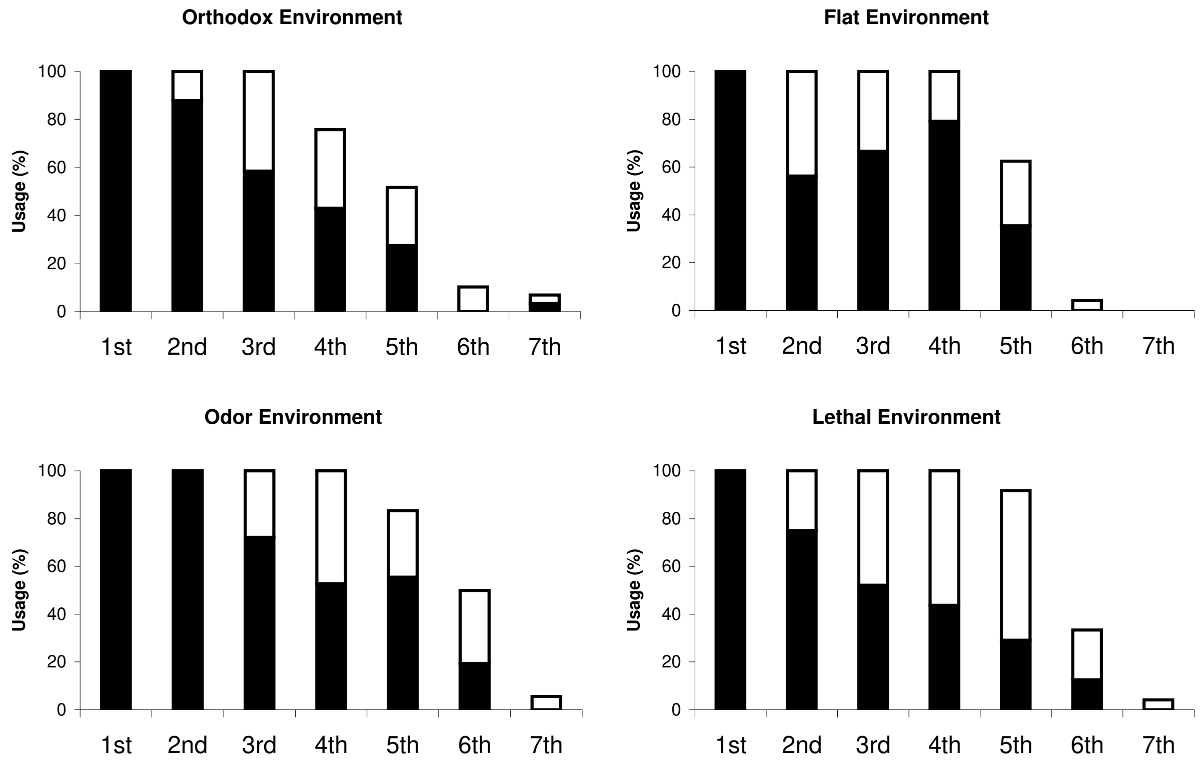

Figure 15: The percentage of elite strategies which treat cues in the predicted (full bars) vs. non-predicted (empty bars) direction across the rank order of cues, for the four environments. Because strategies vary in the number of cues they involve, columns vary in height. Notice that whilst early ranked cues tend to be treated in the predicted direction, the polarity of later cues is not predicted by global validity measures.

reduction in the statistical practice of multiple regression. Many independent variables (cues) may have high predictive power when fitted first, that is, exhibit a high global validity. However, when the complete model (all cues) is fitted, the predictive power of each contributing variable will be less than this first-fitting measure. Discovering the best set of predictors is a problem which cannot be solved by consulting global measures of validity alone.

This problem has implications for lexicographic strategies which order their cues according to global measures of validity, as Take The Best does. Their performance will tend to degrade in increasingly structured environments. This is shown to be true for the Mushroom Problem in Fig. 16, which depicts the mean long-term performance of each set of elite strategies in each environment and the performance of a lexicographic strategy with cues ordered according to their global validities. This Take-The-Best-like strategy does indeed perform adequately in the Orthodox environment, but abysmally in the three structured environments.

In addition, Fig. 16 demonstrates that the ability of strategies evolved within one environment to perform in another varies in an intelligible manner. Whilst the elite strategies evolved within the Orthodox, Flat, and Odor environments perform at essentially the same level within the Orthodox and Flat environments, more of a difference is discernible within the Odor and Lethal environments between "foreign" strategies and those indigenous to the environment. What this demonstrates is that elite strategies from the Orthodox, Flat, and Odor environments can distinguish roughly the same *numbers* of poisonous and edible mushrooms (hence their similar performance in the Orthodox environment); their performance differs in exactly which mushrooms are correctly dealt with and which are incorrectly dealt with (hence their varying performance in the Odor environment). Elite strategies evolved within the Odor environment are less likely to make errors when faced with mushrooms which are significant in their own environment, whereas the errors made by elite Orthodox and Flat strategies are distributed over the space of mushrooms with no concern for their impact in the Odor environment.

One possible explanation for the difference between novices and experts noted earlier stems from these observations. If novices do not appreciate the underlying significance structure of a domain, but experts do, one would expect that in addition to novices perhaps exhibiting a lower level of overall performance, their pattern of successes and errors would not match that of experts, who are more likely to gain their performance from correctly dealing with problems which they consider to be important and/or frequent.

In the Lethal environment the difference between well-adapted strategies and interlopers is most evident. This results from the foreign algorithms' tendency to tolerate a few misses, since their effects can be compensated for by an associated increased number of hits. In the Lethal environment this strategy is clearly maladaptive.

The Take-The-Best-like strategy achieves its low level of performance in the Lethal environment by rejecting every mushroom in favor of the alternative food source. Its conservatism or risk aversion stems from the fact that since no single cue is capable of making error-free recommendations of edibility across the whole space of mushrooms, and errors of this kind are lethal, every cue is best used to reject mushrooms (scoring on average 0.18 rather than negative infinity). As a result every cue is ranked equally and the absence of any cue is taken to be reason enough to reject any mushroom. Since each mushroom will lack at least one cue, every mushroom is eventually rejected by this strategy. (Similarly, within the Flat and Odor environments this strategy uses the presence of any cue as evidence in favor of eating a mushroom, since individually each cue, across the entire population, would best be employed as just such evidence. As a result, all mushrooms are eaten in these two environments and again roughly chance performance is achieved.)

The approach of the elite Lethal strategies falls somewhere between this extreme risk aversion and the blasé attitude to misses exhibited by elite foreign strategies. As discussed above, by using initial cues to exclude particular sets of mostly toxic mushrooms, elite Lethal strategies are able to use subsequent cues to accurately distinguish edible mushrooms from the remainder. In this way they achieve a remarkably competent performance, on average wrongly rejecting (false-alarming) 1 in 10 edible mushrooms and wrongly accepting (missing) no poisonous ones.

4.4 Concluding Thoughts on Significance Structure

Using an artificial foraging task we have demonstrated that manipulating the significance structure of a decision problem can have important implications for the success of decision-making algorithms. We have shown that in order to understand the structure and performance of decision makers

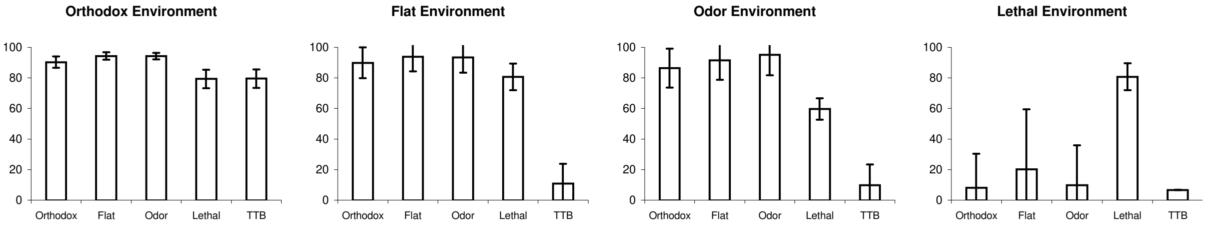

Figure 16: The average mean long-term performance across the four environments of elite strategies evolved for particular environments and a Take-The-Best-like strategy (TTB) that uses cues in order of their global validity. Performance is plotted on the y-axis such that a score of 100 would be obtained by an omniscient and hence perfect forager. In the Orthodox and Lethal environments random behavior would achieve a score of zero. In the Flat and Odor environments, random performance would achieve a score of roughly 10. Whilst the more unstructured environments do not tend to discriminate between groups of elite strategies, the more structured ones favor indigenous elite strategies. TTB performs adequately in the Orthodox environment, but introducing significance structure results in severely reduced performance.

in structured environments an appreciation of this structure is necessary. Significance structure will impact on the performance of strategies in complex ways. Specifically, using global measures of a cue's performance will tend to become misleading as environment structure increases, because the disproportionate contribution of a small number of problem items to a cue's effective performance will cause such global measures of a cue's utility to deviate from the effective utility of a cue within a particular strategy. This was demonstrated for lexicographic cue orderings. Similar lessons are likely to apply to alternative decision heuristics.

5 Overall Conclusions

Rather than conceiving of decision-making success as equivalent to some general-purpose measure of accuracy, the relevant measure is one which captures the extent to which a mechanism copes with its environment, meeting the goals of the decision-making agent. Such a measure must take into account the structure of the agent's environment, including both the environment's frequency structure and its significance structure. Employing this ecologically motivated form of assessment leads to a new vision of what constitutes a good decision making algorithm — sacrificing traditional notions of accuracy and generality can reveal the advantage of heuristics that evidence an increased ability to cope with specific real environments despite their failure to meet internalist criteria of rationality.

Acknowledgements

This paper benefited from the comments of Valerie Chase, Jason Noble, and Henrietta Wilson, discussion with Martin Lages and the ABC Group, and the programming assistance of Martin Dieringer, Torsten Mohrbach, and Rüdiger Sparr.

References

Anderson, J. R. (1991). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Blake, C., Keogh, E., & Merz, C. J. (1998). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`. University of California, Irvine, Dept. of Information and Computer Sciences.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Brooks, R. A. (1991b). New approaches to robotics. *Science*, *253*, 1227–1232.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Cliff, D., Harvey, I., & Husbands, P. (1993). Explorations in evolutionary robotics. *Adaptive Behavior*, *2*(1), 71–108.

Clutton-Brock, T. H., & Albon, S. D. (1979). The roaring of red deer and the evolution of honest advertisement. *Behaviour*, *69*, 145–170.

Connolly, T. (1999). Action as a fast and frugal heuristic. *Minds and Machines*, *??*, ??–??

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In Dupré, J. (Ed.), *The Latest on The Best: Essays on Evolution and Optimization*, pp. 277–306. MIT Press/Bradford Books, Cambridge, MA.

Cummins, D. D., & Allen, C. (Eds.). (1998). *The Evolution of Mind*. Oxford University Press, New York.

Davison, M., & McCarthy, D. (1988). *The Matching Law: A Research Review*. Erlbaum, Hillsdale, NJ.

Dawes, R. M. (1988). *Rational Choice in an Uncertain World*. Harcourt Brace Jovanovich, Orlando, FL.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.

Delius, J. D., & Siemann, M. (1998). Transitive responding in animals and humans: Exaptation rather than adaptation? *Behavioural Processes*, *42*, 107–137.

Evans, J. S. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*, 356–363.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA.

Gigerenzer, G., Todd, P. M., & the ABC Group (1999). *Simple Heuristics that Make Us Smart*. Oxford University Press, New York.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The Take The Best heuristic. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 75–96.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 3–36.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 37–58.

Goodie, A. S., Ortmann, A., Davis, J. N., Bullock, S., & Werner, G. M. (1999). Demons vs. heuristics in artificial intelligence, behavioral ecology, and economics. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 327–356.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*, 517–546.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 209–234.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbour. Reprinted by MIT Press, 1992.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortext. *Journal of Physiology*, *148*, 574–591.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York.

Klauer, K. C. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review*, *106*, 215–222.

Laplace, P. S. (1951). *A Philosophical Essay on Probabilities*. Dover, New York. (F. W. Truscott and F. L. Emory, Trans.; Original work published 1814).

Marr, D. (1982). *Vision*. Freeman, San Francisco, CA.

Martignon, L., & Schmitt, M. (1999). Simplicity and robustness of fast and frugal heuristics. *Minds and Machines*, *??*, ??–??

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. MIT Press/Bradford Books, Cambridge, MA.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT Press/Bradford Books, Cambridge, MA.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press/Bradford Books, Cambridge, MA.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal selection. *Psychological Review*, *101*, 608–631.

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

Schick, F. (1986). Dutch bookies and money pumps. *Journal of Philosophy*, *83*, 112–119.

Schlimmer, J. S. (1987). Concept acquisition through representational adjustment. Tech. rep. 87-19, Deptartment of Information and Computer Science, University of California Irvine.

Shafir, S. (1994). Intransitivity of preferences in honeybees — support for comparative-evaluation of foraging options. *Animal Behaviour*, *48*, 55–67.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acat Psychologica*, *81*, 75–86.

Sober, E. (1994). *From a Biological Point of View: Essays in Evolutionary Philosophy*, chap. The adaptive advantage of learning and *a priori* prejudice, pp. 50–70. Cambridge University Press, Cambridge.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*. Princeton University Press, Princeton, NJ.

Todd, P. M., & Miller, G. F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In *Simple Heuristics that Make Us Smart* (Gigerenzer et al., 1999), pp. 287–308.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choice. *Journal of Experimental Psychology*, *71*, 680–683.

Webb, B. (1996). A robot cricket. *Science*, *275*, 62–67.

Webb, B. (1994). Robotic experiments in cricket phonotaxis. In Cliff, D., Husbands, P., Meyer, J.-A., & Wilson, S. W. (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, pp. 45–54. MIT Press.

Williams, B. A. (1988). Reinforcement, choice, and response strength. In Atkinson, R. C., Herrnstein, R. J., Lindzey, G., & Luce, R. D. (Eds.), *Stevens' Handbook of Experimental Psychology*, pp. 167–244. Wiley, New York.