Reasoning, rationality, and architectural resolution

BRIAN J. SCHOLL

ABSTRACT  Recent evidence suggests that performance on reasoning tasks may reflect the operation of a number of distinct cognitive mechanisms and processes. This paper explores the implications of this view of the mind for the descriptive and normative assessment of reasoning. I suggest that descriptive questions such as "Are we equipped to reason using rule X?" and normative questions such as "Are we rational?" are obsolete—they do not possess a fine enough grain of architectural resolution to accurately characterize the mind. I explore how this general lesson can apply to specific experimental interpretations, and suggest that 'rationality' must be evaluated along a number of importantly distinct dimensions.

Here are two interesting questions: (1) On which kinds of reasoning tasks do we typically succeed, and on which do we typically fail? (2) What exactly is it about the human cognitive architecture that underwrites and explains this pattern of performance? In principle, both sorts of questions are central to the study of human reasoning. In practice, however, the importance of the latter sort of question is not always fully appreciated. Occasionally, the distinction blurs as the results of some experiment are discussed not as data points, but as explanations of the character of cognition. If, for example, it is found that we are often unable to correctly apply some logical or probabilistic rule in a given situation, then it is concluded that the rule is not encoded as a part of our 'mental logic', and we are thus irrational with respect to that rule.

'Reasoning', in this caricatured treatment, amounts to an indivisible cognitive mechanism, and performance in the experimental task is implicitly seen as resulting from a single cognitive process. In such a view of the mind, one can readily reason backwards from faulty performance in a reasoning task to faulty operation of the cognitive mechanism of 'reasoning'. This makes it relatively easy to evaluate reasoning, both descriptively and normatively.

There is some reason to believe, though, that the mind is not so organized. Recent evidence suggests that performance on reasoning tasks may reflect the operation of a number of cognitive processes, many of which may be distinct from

Brian J. Scholl, Rutgers Center for Cognitive Science (RuCCS), Psychology Annex, Busch Campus, Rutgers University, New Brunswick, NJ 08903, USA. Email: scholl@ruccs.rutgers.edu.

0951-5089/97/040451-20 © 1997 Carfax Publishing Ltd
our pretheoretic conception of 'reasoning'. Furthermore, the cognitive processes which do subserve what we pretheoretically call 'reasoning' may be housed in a number of independent (and perhaps domain specific) cognitive modules.

This modern view of the mind has weighty implications for the assessment of human reasoning, which I explore in this paper. When asking and answering questions about such a mind, one must strive to attain as fine a grain of architectural resolution as possible. Descriptive questions such as “Are we equipped to reason using rule X?” and normative questions such as “Are we rational?”—questions which implicitly characterize reasoning as resulting from an indivisible cognitive mechanism—may prove obsolete. Much of the existing discussion on such issues, however, has been implicitly conducted via just these sorts of questions, and may have to be reworked. The goal of this paper is to characterize what such a reworking might involve.

In what follows, I discuss two examples of how attention to architectural resolution forces us to reinterpret certain experimental results, and their implications for the evaluation of reasoning and rationality.

1. Example 1: reasoning and florid modularity

One of the most important results that has emerged from cognitive science thus far is the realization that what we pretheoretically term 'cognition' is actually an amalgam of a number of distinct mechanisms and processes. There are familiar reasons to believe, for example, that the mechanisms which subserve visual object recognition are not the same mechanisms which subserve syntactic parsing, or commonsense reasoning. The mind is not an undifferentiated, indivisible entity.

Researchers from the emerging field of evolutionary psychology have recently taken this insight to heart, with a vengeance. They have championed a view of the mind as floridly modular, overflowing with distinct, innately specified, domain-specific subsystems. They have also marshalled an impressive array of support for this position, including both theoretical arguments as to the adaptive value of different cognitive architectures, and experimental demonstrations of the existence of specific cognitive modules.

1.1 The modularity of mind

I will assume that the notion of modularity, à la Fodor (1983), is familiar, and will only give a quick gloss here. The essence of architectural modularity is a set of restrictions on information flow. The boundaries of modules in this sense are (either one-way or two-way) informational filters: either some of the information inside the module is not accessible outside the module (the so-called 'interlevels' of processing), or some of the information outside the module is not accessible inside the module.

By the first criterion, a central system does not have access to the processing ('interlevels') of the module, but only its output. Thus, for example, if early vision comprises a module (as argued, for example, by Pylyshyn, submitted) whose
output is a certain type of representation (e.g. a ‘primal sketch’; Marr, 1982), then no other extramodular cognitive processes have access to the computations by which the primal sketch is constructed, but only to the sketch itself. By the second criterion, the modularized processes have no access to any external processing or resources. The standard examples here are visual illusions. In the Müller-Lyer illusion, for example, you still fall prey to the mistaken percept that one line is longer than the other, even though you may have convinced yourself (say, by use of a ruler) that they are in fact the same length. This is to be explained by appeal to informational encapsulation: the processes which the module uses to construct your percept are encapsulated with respect to your beliefs. A cognitive function is architecturally modular to the degree that it is informationally ‘encapsulated’ in this way.

These restrictions on information flow engender a number of other ‘symptoms’ of modularity: (a) Modules work in a mandatory way, such that their operation is not entirely under voluntary control. (b) Modules are typically fast, perhaps due in part to the fact that they are encapsulated (needing to consult only a circumscribed database) and mandatory (not needing to waste time determining whether or not to process incoming input). (c) Modules offer highly constrained ‘shallow’ outputs, which themselves often undergo further processing down the line. (d) Modules may often (though need not) be implemented in fixed specialized portions of neural architecture (cf. Scholl, in press). And, (e) modules, and the abilities they support, may be selectively impaired by neurological damage.

The cognitive modules discussed in this paper are also often domain specific (i.e. they process only certain types of inputs), but this need not be the case for all modules: domain specificity refers to the sorts of questions a system can answer, while informational encapsulation refers to the information that the system can make use of whilst answering. You could thus have an unencapsulated (and thus non-modular) domain-specific system (“it answers a relatively narrow range of questions but in doing so it uses whatever it knows”) or an encapsulated but domain-general system (“it will give some answer to any question; but it gives its answers off the top of its head”) (Fodor, 1983, pp. 103–104).

Architectural modules have most of these symptoms, and often exhibit all of them. For a complete discussion of these symptoms, and of modularity in general, see Fodor (1983). For concise summaries of the controversy which currently surrounds modularity, see Carston (1996), Garfield (1994), and Segal (1996).

1.2 Florid modularity

In Fodor’s initial presentation, cognitive modules were identified with (and only with) input systems, which effectively translate transducer outputs into a format which central systems can understand. For Fodor, input systems are comprised of “the perceptual systems plus language” (1983, p. 44). Thus, although there may be many Fodor-sanctioned cognitive modules, they all cluster around visual perception (e.g. for texture segregation or for computing structure from motion) and language (e.g. for phonological segmentation or for syntactic parsing).
Evolutionary psychologists have recently argued that there are far more cognitive modules than we have heretofore suspected. Cosmides and Tooby (1994), for example, suggest that "The human cognitive architecture probably embodies a large number of domain-specific 'grammars', targeting not just the domain of social life, but also disease, botany, tool-making, animal behavior, foraging and many other situations that our hunter-gatherer ancestors had to cope with on a regular basis" (p. 71).

This 'florid modularity' hypothesis is notable in its break—both quantitatively and qualitatively—from initial conceptions of modularity. Whereas Fodor explicitly posits the existence of only a few modules involved in early vision and linguistic processing, evolutionary psychologists suggest that our minds contain "hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestors" (Tooby & Cosmides, 1995, p. xiv; my emphasis). The theory is also notable for the kinds of modules it proposes. Whereas Fodor's modules were always 'input systems'—modules dedicated to the translation of transducer outputs into a language of thought—the most popular modules discussed in the evolutionary psychology literature subserve higher-level tasks—such as reasoning—more closely associated with Fodor's 'central systems'.

How have all these cognitive modules gone unsuspected for so long? Evolutionary psychologists note that the inferences drawn by reasoning modules seem so natural that it is hard to see that there is a problem being solved in the first place. Cosmides and Tooby refer to this situation as instinct blindness:

[T]he 'naturalness' of certain inferences acts to obstruct the discovery of the mechanisms that produced them. Cognitive instincts create problems for cognitive scientists. Precisely because they work so well—because they process information so effortlessly and automatically—we tend to be blind to their existence. Not suspecting they exist, we do not conduct research programs to find them. (Cosmides & Tooby, 1994, p. 67)

One of the corrective lenses for this type of blindness is thinking in terms of evolutionary biology. Evolutionary psychologists thus study the mind not by asking what it is capable of doing, but by asking what it was designed to do. Researchers identify adaptive problems faced by our ancestors, and the information that would have been available in ancestral environments with which to solve them. When one takes this evolutionary perspective, they argue, it becomes clear that the mind simply must be floridly modular.

This line of reasoning is supported by a wide variety of theoretical arguments from the adaptive value of such florid modularity, which I will not explore here (see Cosmides & Tooby, 1992, 1994; Tooby & Cosmides, 1992). Rather, I would like to discuss a set of experimental studies designed to demonstrate the existence of a particular (heretofore unsuspected) cognitive module—a mechanism for reasoning about cheater-detection—and demonstrate how it forces us to radically alter our evaluation of reasoning and rationality in the context of a specific experimental task [1].
1.3 The cheater-detection hypothesis

Rather than attempting to explore the architecture of the mind from the bottom up, with no preconceived expectations, many evolutionary psychologists begin by focusing on adaptive problems that our ancestors faced, and then attempting to demonstrate the existence of the mechanisms which (they argue) must be a part of our cognitive architecture, given that these problems were solved. Cosmides and Tooby, for example, "wanted to ... use an evolutionarily derived computational theory to discover cognitive mechanisms whose existence no one had previously suspected" (1994, p. 60). Their choice: certain specialized domain-specific reasoning mechanisms.

In particular, they focused on certain phenomena of social exchange, which proved important to our ancestors, and which continue to be universally integral to our lives today. One such phenomenon is cheating. Individuals who are able to reap benefits gained from their neighbors (e.g. sharing in a community food supply) without paying the normally-expected cost (e.g. contributing to that food supply) will often enjoy a selective advantage. Correspondingly, it is of paramount importance to the group to be able to detect such individuals and deny them these unearned benefits. Evidence from evolutionary biology (e.g. Axelrod, 1984; Trivers, 1971) suggests that "social exchange cannot evolve in a species unless individuals have some means of detecting individuals who cheat and excluding them from future interactions" (Cosmides & Tooby, 1994, p. 49).

This leads Cosmides and Tooby to posit "that the human mind contains specialized circuits designed for reasoning about adaptive problems posed by ... social exchange" and more particularly "that humans might have evolved inference procedures that are specialized for detecting cheaters" (Cosmides & Tooby, 1994, pp. 60–61)—a 'cheater detection' reasoning module. They offer a number of theoretical arguments (from the nature of design by natural selection) that there must be such a mechanism, but more important for our purposes is an experimental demonstration of its existence, which employed the Wason selection task.

In Wason's selection task (e.g. Wason & Johnson-Laird, 1972), subjects are asked to identify the sorts of information that would falsify a certain claim, presented in the form of a conditional rule. Here's a paradigmatic example. You are presented with the following cards, lying on a table, each of which contains a letter on one side and a number of the other.

\[
\begin{array}{ccc}
D & K & 3 & 7 \\
\end{array}
\]

You are presented with the following claim, and are asked to identify all the cards which need to be turned over in order to determine if the claim is true.

Claim: If a card has a 'D' on one side, then it has a '3' on the other side.

This claim is in the form of a conditional, and the cases relevant to determining its truth are thus the antecedent and the negated consequent—in this case the 'D' and '7' cards.
All manners of subjects are surprisingly poor at this problem. (In one of the original studies, only 5 out of 128 subjects gave the correct answer; Wason & Johnson-Laird, 1972.) The most popular answers are the antecedent and the consequent (i.e. the ‘D’ and ‘3’ cards in our example: the two cards explicitly mentioned in the claim itself), followed by the antecedent alone. A long-standing interpretation of this pattern of results is that we tend to look for information that will verify a rule, but not for information which will falsify it. And some writers have taken such faulty performance, along with this bias, as evidence that we do not naturally reason according to the correct principles of logic, and are thus irrational to a degree. The experiments of Cosmides (1989), however, seriously complicate this picture [2].

Cosmides chose to test out her ‘cheater-detection’ hypothesis using the selection task. She reasoned that if we possessed specialized mechanisms for reasoning about cheater-detection, then those mechanisms would be activated by selection-task problems involving such situations, and performance would improve dramatically. Now consider a paradigmatic (albeit generic) example of this second type of selection task (from Cosmides & Tooby, 1992). You are presented with the following four cards and instructions.

It is your job to enforce the the following law: “If you take the benefit, then you must pay the cost”. The cards below have information about four people. Each card represents one person. One side of a card tells whether a person accepted the benefit and the other side of the card tells whether that person paid the cost. Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking this law.

<table>
<thead>
<tr>
<th>benefit</th>
<th>benefit</th>
<th>cost</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>accepted</td>
<td>not accepted</td>
<td>paid</td>
<td>not paid</td>
</tr>
</tbody>
</table>

In these ‘cheater-detection’ cases, in marked contrast to the original experiments (e.g. our ‘3 if D’ example), nearly all subjects recovered the (formally identical) correct answers with no appreciable effort at all. In the context of detecting cheaters, the correct answers to such problems just ‘pop out’ and seem not to require any effortful reasoning. In general: “Whenever the content of a problem asks subjects to look for cheaters on a social exchange—even when the situation described is culturally unfamiliar and even bizarre—subjects experience the problem as simple to solve, and their performance jumps dramatically” (Cosmides & Tooby, 1994, p. 63) [3].

1.4 The lesson: architectural resolution

Cosmides and Tooby and their colleagues argue that such experiments demonstrate the existence of a specialized cognitive mechanism for reasoning about cheater detection (Cosmides, 1989; Cosmides & Tooby, 1992, 1994; Gigerenzer & Hug, 1992). Furthermore, they consider this as merely the first demonstration of such
mechanisms, and quite explicitly expect to discover many other distinct domain-specific reasoning mechanisms (e.g. for reasoning about threats and precautions).

As noted above, this view breaks sharply with traditional views of the cognitive architecture of reasoning, which were limited to an ill-defined ‘reasoning box’ which perhaps contained a ‘mental logic’. Many traditional discussions of reasoning and rationality have been conducted in a framework which implicitly assumes that reasoning is subserved by a single, indivisible cognitive mechanism, which (quite simply) either does or does not embody correct principles of inference. If it does, then we are rational creatures; otherwise we are not. If evolutionary psychologists are on the right track with the florid modularity hypothesis, this implicit picture of the cognitive architecture underlying reasoning—and the evaluative commentary which rests on it—must be discarded.

In terms of the descriptive evaluation of reasoning, the lesson is this: Evaluative questions and conclusions must themselves encompass florid modularity in order to attain a fine enough grain of architectural resolution to accurately characterize the mind. Questions such as “Do humans reason according to rule X?” (where X is a law of probability or logic) implicitly treat ‘reasoning’ as an indivisible cognitive process—which as we have seen may not be the case. Arguments as to the right answer to such questions are thus useless. One can give no definitive answer to such a question except: “Yes and no.” ‘Yes’ in some contexts (e.g. cheater detection); ‘No’ in others (e.g. with abstract formal symbols).

In a floridly modular mind, certain logical rules may be implemented in some reasoning mechanisms but not others, and may only be activated by certain types of input. Different reasoning mechanisms may even employ mutually inconsistent rules. The demonstration of faulty performance on some task involving a rule of logic or probability (e.g. Wason’s selection task) therefore does not imply that no cognitive mechanisms operate in accord with that rule! Performance on reasoning tasks may be a clue not to the content of the ‘reasoning box’, but only to the content of a reasoning box, one among many. We would be far better off asking such questions which themselves encompassed florid modularity from the start, such as “Do humans reasoning according to rule X in situations Y?”, where Y is perhaps cashed out in terms of different sorts of subject-domains, and the types of input which characterize the reasoning problem.

The same lessons hold for the evaluation of ‘rationality’. The normative assessment of what we pretheoretically term ‘reasoning’ may in fact be the normative assessment of a number of processes and mechanisms, all importantly distinct. Questions such as “Are we rational creatures?” may have to give way to questions such as “Do we reason rationally according to rules X when reasoning about domains Y?”, which have a finer grain of resolution vis-à-vis the underlying cognitive architecture.

In a floridly modular mind, certain cognitive reasoning mechanisms may operate ‘rationally’, while others do not. Although the charge of “Irrationality!” is certainly warranted in one respect—after all, subjects in these experiments are performing systematically poorly!—this broad conclusion lacks the resolution to encompass what may be the true underlying architectural picture. In certain cases
we may have a situation of global irrationality, in that our performance is systematically faulty, but also of local rationality, in that certain (other) cognitive mechanisms which subserve bona fide reasoning are capable of performing perfectly rationally in such cases [4].

2. Example 2: garbage in, garbage out

The conclusion of the previous section was that certain evaluative questions about reasoning and rationality may be rendered obsolete if the new 'floridly modular' view of the cognitive architecture underlying reasoning is on the right track. In this section I argue for a similar conclusion, but this time in a way which does not depend on there being multiple distinct cognitive mechanisms which underlie 'reasoning'.

Even if there is only a single 'reasoning box', it does not function in complete isolation, but is only one link in a chain of many cognitive mechanisms which together subserve subjects' performances on reasoning tasks. Faulty performance on such tasks may thus fail to justify cries of 'Irrationality!' for another reason: in some cases, the architectural locus of the faulty performance may lie upstream of anything we want to call 'reasoning' (e.g. in encoding mechanisms). This is another situation of global irrationality, in that our performance is systematically faulty, but of local rationality, in that those cognitive mechanisms which actually subserve bona fide reasoning are performing perfectly rationally.

2.1 The 'GIGO' principle

On one of my first forays into computer programming, I had the following frustrating experience. I was trying to code up a conceptually simple but mathematically intense program for graphing certain patterns of complex numbers, and my program stubbornly refused to work: it compiled and ran successfully, but never managed to produce the correct graphs. After debugging for ages, to no avail, I foolishly began to suspect that there might actually be a flaw in my compiler, or perhaps even in the hardware itself.

Of course, it turned out that all the hardware and software was functioning flawlessly, and I had simply failed to specify the correct functions in my code. I had fallen prey to the GIGO principle: Garbage In, Garbage Out. In one sense, the computer was performing poorly—after all, it wasn't giving me the output graphs I required. In another very important sense, though, the computer was functioning impeccably, and the source of the faulty graphs lay not with the computer but with my failure to deliver the correct functions in the first place. Upon encountering the faulty output, I immediately implicated the processing, whereas the true problem was to be found in my faulty input.

The importance of the GIGO principle stems from the ubiquity of the temptation to implicate processing as the source of faulty output. But its validity is really just a simple matter of logic: in a simple system consisting of a black-box processor
which generates certain outputs for given inputs, the cause of ‘faulty’ output may lie with either the black box itself, or with the inputs it receives.

The point of this section is that the same point holds for a ‘reasoning box’. As noted above, writers on reasoning and rationality have sometimes interpreted the results of certain experiments as demonstrating that humans are systematically irrational (see Lopes, 1991 for discussion). Some of these ‘bleak implications’ seem to center on an inference from faulty output (i.e. from poor performance on a reasoning task) to faulty processing (e.g. to a ‘mental logic’ which does not conform to the ‘correct’ principles of logic and probability). But in accord with the GIGO principle, this inference cannot go through without also considering another culprit: faulty input.

A paradigmatic GIGO interpretation, in this context, goes something like this: In the reasoning task, subjects are indeed failing to produce the correct responses. But this may be because crucial aspects of the experimental situation aren’t encoded in the first place, in one important sense or another. The mechanisms which subserve reasoning about the task (the ‘reasoning box’) will then operate on this inadequate input in impeccably rational ways, but will nevertheless produce output which we deem faulty. The reasoning mechanisms give us garbage as output, but only because they got garbage as input in the first place. Even when the input is carefully presented by an experimenter, the cognitive mechanisms which subserve the subjects’ perception and encoding of this input may be the locus of the eventual faulty performance [5].

I turn now to a brief example of how the GIGO principle could apply in an actual experimental interpretation.

2.2 Illusory correlation

When an observer reports the existence of a correlation between two classes of objects or events, she may often be mistaken—the correlation may not in fact exist. Such systematic errors have been termed illusory correlations. More specifically, a reported correlation is illusory just in case the relevant instances (a) are not actually correlated in that sample of data; (b) are correlated, but to a lesser degree than that reported; or (c) are negatively correlated, i.e. in the opposite direction from that reported (Chapman, 1967). This situation will often obtain throughout a population, with many independent observers each reporting the existence of a correlation which is in fact illusory. The study of this phenomenon is of considerable ecological interest, since illusory correlations are ubiquitous in many familiar contexts, some of which are relatively harmless (as in many superstitions, from rain dances to rabbits’ feet), and some of which look to be not-so-harmless, as in the case of clinical diagnosis: many clinicians persist in reporting what appear to be illusory associations between particular clinical disorders and responses on certain projective tests.

Chapman and Chapman (1969) demonstrated the existence of illusory correlation in the clinical domain via the use of Rorschach ink-blot interpretations as predictors of male homosexuality. Many clinicians have reported that male homo-
sexuality is correlated with certain classes of Rorschach interpretations, involving various face-valid predictors, such as ambiguous or incongruent genitalia, anal content, and the like. As it turns out, there are some valid correlations in such data. The problem is that the frequently reported correlations are all illusory, while the valid correlations are never reported [6]! The Chapmans set out to explain why so many clinicians would persist in reporting observations of these illusory correlations in their clinical practices, in the face of a wealth of invalidating empirical evidence.

They first found that reports of semantic association obtained from naive undergraduates correlated almost perfectly with those correlations observed by clinicians in their practice (obtained via anonymous questionnaires). That is, the clinicians were reporting exactly those correlations which naive students judged to be related. Chapman and Chapman (1969) then showed a different group of naive subjects a series of Rorschach cards, each of which included a reported interpretation, and two diagnoses of the hypothetical patient. The Chapmans carefully constructed these cards such that there were no valid correlations between ink-blot interpretations and diagnoses. Nevertheless, the naive subjects reported observing illusory but face-valid correlations—exactly those which had been judged to reflect semantic association, and had been reported by practicing clinicians [7]!

These studies strongly suggest that although plenty of clinicians report observing face-valid correlations in their clinical practice, they are mistaken: the empirical evidence is correct, and the reported correlations are all illusory.

### 2.3 A GIGO interpretation of illusory correlation

Illusory correlation is one phenomenon which has been widely discussed in the experimental reasoning literature (e.g. Nisbett & Ross, 1980; Sutherland, 1992), and which has been taken by some to cast doubt on human rationality. The gist of this move is fairly intuitive and seemingly harmless: since subjects perform systematically poorly on tasks involving judgments of covariation, they must be reasoning irrationally to some degree. It seems clear, however, that this faulty performance need not reflect faulty reasoning at all, à la the GIGO principle.

#### 2.3.1 The assimilation bias

A bit of background first. In cognitive psychology, the term schema is used to refer to "a data structure for representing the generic concepts stored in memory" (Rumelhart, 1980, p. 34). Schemata represent all manners of knowledge in this way—including concepts of objects, situations, and events—and represent it at all levels of abstraction, often embedding within each other. A schema is thus a unit of human knowledge representation that is abstract and structured in certain important ways. Schemata are essentially active processes, which guide encoding and recognition: their central function is to interpret objects, situations, and events in the world. It is also emphasized that schemata are thought to be (at least partially) involved in the perception and encoding of information in a way that is very data driven, and does not necessarily involve deliberate high-level reasoning [8].
“Because you never get a second chance to make a first impression.” This sentiment (a recent marketing slogan for dandruff shampoo, I believe) turns out to be true in a very important way. Here’s the idea: it is much easier for incoming information to activate an already existing schema than it is to create an entirely new schema. Thus, incoming information will be interpreted in the context of pre-existing schemas (“first impressions”) whenever possible. And since phenomenal perception is often mediated by schema activation, that information which does activate schemas will tend to be both encoded and noticed, more so than that information which does not activate any pre-existing schemas.

This can lead to a ‘confirmation bias’ in the actual encoding stage: (a) theories or stereotypes which you hold will tend to be represented in terms of schemas, and any incoming information which activates those schemas will tend to be encoded and noticed; meanwhile, (b) theories or stereotypes which you don’t hold will tend not to have associated schemas, and incoming information supporting these perspectives will fail to ‘find a home’ in a pre-existing schema, and will thus tend not to be encoded or noticed (unless it’s overwhelmingly salient and unique, in which case it may engender the creation of a new schema). When you look back on some set of information, then, you may be biased towards remembering only those data which support a pre-existing theory or stereotype—not through any fault of memory retrieval processes, but because the correct distribution of information was not accurately encoded in the first place. In other words, our encoding mechanisms are positively biased towards that information which can be readily assimilated to our existing schematic model of the world. And this assimilation bias can sometimes get in the way of accurate perception [9].

A number of experimental demonstrations support this picture, two of which I will mention here. The first study emphasizes the effects of this assimilation bias on actual perceptual encoding and recognition. Bruner and Potter (1964) showed subjects slides of familiar objects, which were initially out of focus, and which were brought slowly into focus by discrete steps. At each step, subjects had to give their best guess as to the identity of the object. In this situation, subjects continued to misidentify the object long after the step at which a separate group of subjects—who started with much less severe defocusing—could readily identify it. This experiment demonstrates “the debilitating effect of premature commitment to a particular schema” (Rumelhart, 1980, p. 47).

The second study demonstrates a similar assimilation bias, by manipulating the ease with which a body of information can be assimilated into an existing schema. Bransford and Johnson (1973) presented the following passage to a number of subjects, who were to try and remember as much of it as possible:

The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but
complications can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually, they will be used once more and the whole cycle will then have to be repeated.

However, that is a part of life. (Bransford & Johnson, 1973, pp. 400–401)

The two groups of subjects who heard this passage were treated identically except that one of them was told beforehand that the paragraph they heard would be about washing clothes. And, perhaps unsurprisingly, the group given this context radically outperformed the ‘no-context’ group in terms of comprehension and recall. This too can be explained by appeal to the assimilation bias: the context allowed subjects to assimilate the paragraph in terms of their pre-existing ‘washing-clothes’ schema, facilitating encoding (and thus comprehension and recall) by binding various parts of the story to various properties and variables of the original schema. Note that this is not merely an effect of facilitated retrieval, but of actual encoding: subjects who were given the ‘clothes-washing’ context after hearing the passage did no better on the recall test than those subjects who had never heard the context at all.

In each of these cases, the assimilation bias imposed by schemata is simply a result of the way in which they organize our model of the world: “(B)ecause it predicts that the information picked up (and retained) about new instances is guided by information abstracted from previous instances, the schema view imposes a measure of systematicity on which instances or details are retained” (Komatsu, 1992, p. 12). While doing this, “schemata not only contribute toward the development of an accurate percept, but, by the same token, they can sometimes cause a distortion” (Rumelhart, 1980, p. 47).

2.3.2 The assimilation bias as an application of the GIGO principle. It is just this distortion which may be the true culprit in the illusory correlation experiments. As mentioned above, these experiments are sometimes interpreted as demonstrating faulty reasoning at work—yet another example of human irrationality. Our discussion of the assimilation bias, though, suggests that such results are quite consistent with another sort of interpretation. To wit: garbage in, garbage out.

Subjects in the covariation-judgement tasks, for example—both undergraduates and clinicians—approach their task with various schemas already in place. They already know, for example, about certain stereotypical associations involved with male homosexuality—i.e. all the face-valid predictors in the Rorschach-card experiment. These schemas are instantiated when the observer encounters an appropriate Rorschach interpretation/diagnosis pair, and the observers tend to encode, notice, and thus report that relationship. Again, however, those valid pairs which cannot be assimilated into any existing schemas tend to ‘fall on deaf ears’, as it were, and are not noticed, encoded, or reported [10].

Given that the illicit filtering, on this story, occurred because of the assimilation
bias at the level of schema encoding—mechanisms which plausibly lie upstream of the mechanisms subserving reasoning—subjects' incorrect judgments may not reflect faulty reasoning at all, but (merely?) faulty encoding. Garbage in, garbage out. The clinicians persist in reporting the illusory correlations because in a very real sense they persist in seeing the illusory correlations [11].

2.4 The lesson (again): architectural resolution

This sort of interpretation is actually not so uncommon, I think, but plenty of writers have discussed illusory correlation in a way which suggests that it is somehow an effect of reasoning, and thus may impact our assessment of rationality. Sutherland (1992), for instance, cites the Chapmans' research as demonstrating "the irrationality of human judgement" (p. 161). More innocently, here are Nisbett and Ross: "Beliefs about covariation may be formed through strategies which do not meet conventional logical or statistical criteria but which provide reasonably accurate estimates under some circumstances" (1980, p. 109; my emphases). This illustrates a common supposedly 'safe' interpretation of poor reasoning performance: 'at a minimum, it is certainly the case that mechanisms of belief fixation have performed sub-optimally'.

This is quite true, but let such statements not obscure the fact that there are multiple distinct mechanisms of belief fixation (e.g. via reasoning and via perception), and that a failure of one such method may not have the same sorts of implications as does a failure of another. In the case of illusory correlation, the GIGO interpretation suggests that there is an important distinction between (a) failing to reason appropriately about information you have successfully encoded, and (b) failing to reason appropriately about information you haven't successfully encoded. The GIGO interpretation suggests that (b) might be appropriate in some situations, in which case the status of reasoning mechanisms and of human rationality may not be implicated at all.

As in the previous example, the charge of "Irrationality!" is certainly warranted in one respect—after all, subjects in these experiments are again performing systematically poorly. But again this broad conclusion lacks the resolution to encompass what may be the true underlying architectural picture. The GIGO story suggests that in certain cases we may have a situation of global irrationality, in that our performance is systematically faulty, but also of local rationality, in that those cognitive mechanisms which actually subserve bona fide reasoning are performing perfectly rationally.

3. Conclusions: global irrationality, local rationality, and architectural resolution

We've now seen the same point made two ways. In the contexts of both a floridly modular mind and a single reasoning mechanism, we've seen that it is much more difficult to determine the locus of faulty task performance—and thus to draw sweeping conclusions about reasoning and rationality—when you view the mind with a fine enough grain of architectural resolution.
To make these points as intuitive as possible, consider the three situations depicted in Figure 1. In the first case, all the processing takes place in a single indivisible 'reasoning' mechanism. In the second case, the output could have been produced by any one of the reasoning mechanisms, arrayed in parallel. In the third case, the output was produced by a stream of different processing mechanisms, arrayed in serial.

Exercise: based on the observation of faulty performance in each case, what can we conclude about the highlighted mechanism in terms of reasoning and rationality?

In the first 'indivisible reasoning' case, the conclusion is obvious. Since (a) the performance was systematically faulty, and (b) only a single cognitive process ('reasoning') subserved that performance, it must be the locus of the fault. On such a model of the mind, one can (all too) readily draw the corresponding normative conclusions: the results indicate that the reasoning mechanism does not behave rationally.

In the second 'florid modularity' case, the conclusion is also obvious. Since you
have no way of knowing (without further investigation) which reasoning mechanism was actually involved in producing the faulty performance you observed, you can conclude absolutely nothing about the highlighted mechanism, since it may not have been involved at all. You may draw conclusions about the disjunction of all the mechanisms ('global rationality'), but not about any particular distinct mechanism ('local rationality').

The same holds in the third GIGO case. Since a number of distinct cognitive mechanisms were involved in the task performance, you have no way of knowing which one introduces the fault. So again, you can conclude nothing at all about the highlighted mechanism, but only about the disjunction of the system as a whole.

And of course, these last two cases are not mutually exclusive. The arguments above suggest that the cognitive architecture of reasoning may be populated by multiple distinct cognitive mechanisms arrayed in both serial and parallel, making it all the harder to pinpoint the guilty party.

3.1 Global irrationality, local rationality, and the candidates for normative assessment

For these reasons, I have argued that 'rationality' may be an obsolete concept, and that finer distinctions are needed. There is an unambiguous target for the normative assessment of rationality only when your implicit model of the mind contains a single indivisible cognitive mechanism which suberves performance on a reasoning task. And as we have seen, there are many ways in which such a model could be false. I have suggested that in order to normatively evaluate the cognitive architecture underlying reasoning, we will need to ask normative questions which themselves possess a finer grain of architectural resolution.

Of course, we will still want to distinguish between those cases in which subjects perform reasoning tasks correctly, and those cases in which they do not. Defining 'rationality' in terms of this distinction would be perfectly acceptable, but 'rationality' would then cease to be a very interesting focus for normative evaluation. I have suggested above that we might more usefully distinguish between global rationality and local rationality. The former term can apply to the distinction just noted, between correct and faulty performance on reasoning tasks resulting from the operation of the cognitive architecture as a whole, while the latter can refer to the normative evaluation of specific cognitive mechanisms, in the typical ways.

In a floridly modular mind, we may often have situations of global irrationality but local rationality: the substandard logic of one reasoning mechanism does not preclude there being many other reasoning mechanisms which operate flawlessly in this regard. And the same conclusion holds in a non-floridly modular mind, à la the GIGO principle. We may often have situations of global irrationality—i.e. of faulty performance on reasoning tasks—but of local rationality, in that the locus of the faulty performance may lie in cognitive mechanisms upstream from the reasoning mechanism, which is itself performing impeccably rationally.

Here's a bad question: "Are we rational?" Here's a better (but still bad) question: "Are our type-X mechanisms rational?" (For example: mechanisms for encoding, response selection, and 'reasoning'.) And here's a better question: "Are
our type-X mechanisms rational with respect to domain Y?" Only this last sort of question can encompass the underlying cognitive architecture suggested by the florid modularity hypothesis and the GIGO principle.

A precise normative evaluation of performance on reasoning tasks, of course, will likely require us to distinguish many different kinds of local rationality. To conclude, I'd like to identify some of these candidates for normative assessment.

1. In many situations, we will want to evaluate the performance of both ‘reasoning’ mechanisms and the upstream ‘encoding’ mechanisms. For instance, we may very well wish to draw ‘bleak implications’ for encoding mechanisms by subjecting the ‘assimilation bias’ discussed in Section 2.3 to normative evaluation itself (but cf. note 11). We might call this ‘encoding rationality’.

2. In order to evaluate a floridly modular mind, we may wish to evaluate our performance on reasoning tasks which span a number of distinct domains. This would correspond to the evaluation of a domain-specific reasoning module on its ‘preferred input’.

3. We may find it useful to evaluate the mind as a whole in terms of what kinds of domain-specific reasoning mechanisms it contains. If we have a mechanism for reasoning about cheater-detection, but not a mechanism for reasoning about the stock market, does this mean that we are irrational with regard to this latter domain? Or does it merely mean that such a domain is not adaptively important?

4. It is still an open question just how the domain-specificity of certain reasoning mechanisms is implemented. It is possible (a) that incoming input (e.g. from a reasoning problem) is ‘broadcast’ to all the reasoning modules, each of which only takes action given certain input; or (b) that a separate mechanism ‘routes’ various sorts of input to the appropriate modules. In this latter control structure, we may be able to evaluate the efficacy of this ‘routing’ mechanism [12]. Does a module always receive all the sorts of input on which it can engender successful performance, or might there be cases where a certain module could reason about a certain situation successfully, but does not, simply because it never gets ‘sent’ the input in the first place? Is this a suboptimal situation (‘routing irrationality’?), or merely a necessary consequence of an efficient implementation of domain-specificity?

5. Finally, how shall we evaluate the situation wherein one cognitive mechanism compensates for the local irrationality of another? For instance, suppose that I (deliberately and consciously) employ the following heuristic: (1) whenever presented with a selection-task problem, I construct a mapping between the given situation and one which involves cheater-detection; (2) I then reason about the cheater-detection version, coming up with an answer; and (3) I then follow the mapping back to recover the correct answer for the given problem. According to the hypotheses of Cosmides and Tooby discussed above, this ought to result in successful performance on a whole range of selection tasks. This would be the inverse of the cases discussed above: I am globally rational—in that I’m performing successfully on the given selection tasks—but locally irrational—in the sense that the cognitive mechanisms which would normally subserve reasoning about
such problems (before I started employing my heuristic) do not contain the normatively appropriate rules.

3.2 A concluding prescription: practice architectural resolution!

These are the sorts of candidates for normative assessment with which a theory of 'rationality' must contend. Each of them reflects different aspects of a complex underlying cognitive architecture, and neither has any a priori precedence over the others. 'Rationality' is an outdated concept, which must be replaced with several different 'rationalities'.

The lesson? Practice architectural resolution! Don't draw global conclusions (about 'reasoning' or 'rationality') without explicitly pointing to either (a) the local architectural mechanisms you think are responsible, or (b) the fact that you're remaining agnostic on their identities. Architectural resolution should always be present, even when it's only a statement of ignorance.

Much recent work in the cognitive psychology of reasoning has involved ferreting out these puzzles on which we seem to perform poorly. Occasionally, such results can seem to be characterized as explanations of the character of cognition. In this caricatured view, poor performance on a reasoning task involving some rule of logic or probability indicates that humans do not reason according that rule, and are thus irrational. But a characterization of our performance on one of these puzzles is not a final explanation, but a data point, crying out for architectural interpretation.

I have tried to demonstrate here that there is perhaps a wider range of possible architectural interpretations of this research than has been generally realized, and that this complexity may have weighty implications for how we ought to evaluate human reasoning and rationality.

Acknowledgments

I would like to thank Gretchen Chapman, Zenon Pylyshyn, Richard Samuels, Polly Tremoulet, and especially Stephen Stich for stimulating conversation, and for helpful comments on earlier drafts. Thanks also to two anonymous reviewers for several helpful suggestions.

Notes

[1] In fact, I do think that some serious problems infest the theoretical arguments used by Cosmides and Tooby in support of these conclusions. My purpose in this paper is not to evaluate the 'florid modularity' hypothesis, however, but to explore its implications for the proper evaluation of reasoning and rationality.

[2] Actually, this picture was seriously complicated much earlier. It was quickly noticed that performance on the selection task improves markedly when the context employs ecologically familiar rules rather than abstract symbols (e.g. Griggs & Cox, 1982), but the preferred explanations of this improvement—in terms of familiarity and concreteness—have been called into question by the evidence discussed here.

[3] For many more details concerning these experiments, and the failure of various alternative
hypotheses to account for their results—e.g., explanations based on familiarity, availability, concreteness, permission schemas, payoff possibilities, and mere facilitation of domain-general reasoning—see Cosmides (1989), Cosmides and Tooby (1992), and Gigerenzer and Hug (1992). High levels of correct performance are only observed in situations where the subject is cued into the perspective of somebody who is motivated to detect cheaters. For a recent dissenting interpretation, see Liberman and Klar (1996).

[4] More on this distinction in Section 3. I am indebted to Richard Samuels for first framing the distinction in these terms.

[5] This GIGO story depends on a certain picture of the cognitive architecture, in which the mechanisms that subserve the encoding of environmental stimuli are in some sense distinct from—and lie upstream of—those which subserve 'reasoning'. This strikes me as an eminently plausible (even boring) assumption, tantamount to the claim that the mechanisms which subserve reasoning do not have immediate access to the world, but can only operate on input which itself is the result of previous distinct encoding processes. In any case, a full-dress defense of this picture is far beyond the scope of this paper, and I shall simply assume it in the following.

[6] The two valid predictors of male homosexuality are (a) responses to Card IV involving a human or animal which is somehow contorted, monstrous, or threatening, and (b) responses to Card V involving a human animal (e.g. a woman dressed as a bat) or a humanized animal (e.g. a centaur). Whereas there are fairly obvious explanations as to why the illusory correlations may obtain, it is far from clear why these particular correlations are valid.

[7] Further manipulations suggested that valid (but non-face-valid) correlations are occasionally reported accurately to some degree, but that this does not affect the report of illusory correlations: even perfect non-face-valid correlations are reported as weaker than those which are completely invalid but face-valid. Furthermore, these illusory correlations do not disappear over time with practice, and are not affected by monetary rewards for accuracy (ruling out a simple account in terms of motivation). Only when subjects are allowed to sort and repeat cards while taking notes are the illusory correlations attenuated (Chapman & Chapman, 1967, Experiment VI).


[9] See Evans and Over (1996) for a similar thesis, wherein "explicit or conscious thinking is focused on highly selected representations which appear 'relevant' but that this relevance is determined mostly by preconscious and tacit processes" (p. 48). For a review of related evidence and critical discussion, see Alba and Hasher (1983), who discuss and evaluate evidence that "What is encoded, or stored in memory is heavily determined by a guiding schema or knowledge framework that selects and actively modifies experience in order to arrive at a coherent, unified, expectation-confirming and knowledge consistent representation of an experience" (p. 203).

[10] Note that in the present paper, I'm not claiming that this bit of speculation is necessarily correct, but merely possible. My purpose here is not to argue for any specific interpretation of any specific set of experiments, but rather to focus on the variety of possible interpretations that exist once we frame our questions with a fine enough grain of architectural resolution.

[11] One might ask why this would be the case, from an adaptive standpoint. Why would we have evolved schemas which can give rise to this assimilation bias, and thus to faulty encoding? Surely it is of some adaptive value to be able to accurately note correlations in the environment. I have not addressed these sorts of adaptive considerations in this paper, but it seems plausible that this situation reflects a compromise of sorts, between accurate correlation detection and efficient cognitive processing. The assimilation bias may be the price paid for a system which can function at all in the real world. Evans and Over (1996) discuss this sort of argument in the context of their related theory: "Positive information is preconsciously selected as relevant because it is effective to do so given the limitations on human information processing, working memory capacity, and the time available for searching the world for evidence" (p. 106). See Evans and Over (1996) for a full defense of this view, in which the (sometimes normatively faulty) operation of various cognitive processes involved in reasoning tasks is taken to "reflect processes that are adaptive in the real world if not in the laboratory" (p. 49). (Thanks to an anonymous reviewer for raising this concern.)
RATIONALITY AND ARCHITECTURAL RESOLUTION 469

[12] Cf. Pylyshyn’s two distinctions: “(1) between sending control (where the initiative lies with the old locus) and capturing control (where the initiative lies with the new locus), and (2) between directing a message to one specified recipient and broadcasting it to all routines or ‘modules’ at once” (1989, p. 68).

References


COSMIDES, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason?


