Is psychology hard or impossible? Reflections on the conditional.

Keith Stenning Michiel van Lambalgen

May 25, 1999

Abstract

We discuss the various explanations that have been offered to account for subjects' behaviour in Wason's famous selection task, and find them wanting. We argue that what is lacking is a good understanding of a subject's semantics for the key expressions involved, and an understanding of how this semantics is affected by the demands the task puts upon the subject's cognitive system.

1 Introduction

When Peter Wason invented his '4-card' task (e.g. Wason [28]), he created one of cognitive science's fruit flies—a laboratory phenomenon of deceptive simplicity which was a potential basis for theory which could reach far beyond its confines. The purpose of this paper is to review the extent to which that promise has thus far been fulfilled. Our argument will be that this topic has the potential to unite disparate areas of cognitive science, but that existing explanations do not make much attempt to do so. We sketch one possible integration of accounts of the semantics of the conditional with the existing behavioural evidence.

2 Task and Phenomena

Wason's task involves the choice of evidence in support of a conditional rule. The reasoner is presented with four cards, and told that each has a letter on one side and a number on the other. A conditional rule describes the four cards. In Wason's original experiment this rule was "If there is a vowel on one side of the card, then there is an even number on the other". The reasoner's task is to turn those cards and only those cards which it is necessary to turn in order to see if the rule is true. Four cards bearing, say, 'A', 'K', '4' and '7' appear below the rule.

In this and many subsequent replications, intelligent undergraduate student populations have shown a range of card choices, but very few students produce the normative response of choosing the cards which exhibit the true antecedent and false consequent on their visible faces (A and 7 in the example above). The modal response is to choose the true antecedent and true consequent cards.

Almost all students choose to turn the A. Many turn the 4. Some turn the K. And very few turn the 7.

The distribution of choices from Wason and Johnson-Laird's [29] table is: p, q 46%; p 33%; $p, q, \neg q$ 7%; $p, \neg q$ 4%, and others 10%.

Very similar data have been obtained many times. More importantly, the experiment has been run with many variations, particularly of rule content and task instructions, and much is known of what is observed in these various circumstances. Wason's task is known in the conditional reasoning literature as the *selection* task to distinguish it from several other widely used tasks, notably the *evaluation* and *construction* tasks which have also been applied to the study of conditionals. The evaluation task presents a conditional rule, and a particular 'case' (in terms of values for antecedent and consequent) and asks whether the rule is true of the case. The construction task presents a rule and asks subjects to construct a case of which the rule is true, and one of which it is false.

So, much is known about the behavioural facts of conditional reasoning, and one might hope that this contribution of the psychology of reasoning would be of obvious relevance to a number of other communities of researchers—logicians, philosophers of science and language, linguists, those interested in normative theories of induction, decision making and machine reasoning. The study of conditionals, has, after all, been a major concern of philosophers and semanticists. Symmetrically, one might suppose that what is known about the semantics and pragmatics of the conditional might be frequently drawn upon by the psychologists concerned with explaining what is observed in the selection task. One might even suppose that those concerned with the education of undergraduate students in the arts of reasoning and communicating might have some interest in this set of, at least apparently, scandalous observations.

Instead the situation is rather different. It is true that Wason made a connection right from the outset with Popper's philosophy of science. Indeed, Popper's philosophy seems to have played a central role in inspiring Wason's invention. We will see below how this figures in some of the explanations given for some of the phenomena. But there is virtually no contact between psychologists working in this tradition and those studying the semantics of conditionals or the nature of rules and laws. Fillenbaum's [9] work is a worthy exception, but perhaps one that proves the rule. There has been some linguistic interest (Geis and Zwicky [11]) in the relation between the psychological observations and the theory of pragmatics. Philosophical work on the Ravens Paradox (see below) has been cited in support of statistical theories of students' reasoning. But by and large, the theories of performance in these tasks has not been related to what other disciplines have contributed to the understanding of conditionals.

One reason for this is that several of the psychologists involved have seen these observations as knock-down arguments against the employment of formal theories in explaining students' behaviour (e.g. Johnson-Laird and Wason [30]; Johnson-Laird and Byrne [16]). This response has especially been engendered by what are known as *thematic* or *content* effects. Early after Wason's initial experiment, Wason and Shapiro [31]) and Wason and Johnson-Laird [30] experimented with conditional rules which, in context, made the connection between antecedent and consequent more vivid: *If I go to Manchester, I go by train* and *If the envelope is sealed, it must have a first class stamp* respectively. Such material has come to be known as *thematic* as opposed to the *abstract* letters and numbers of the classical experiment. Of course, the letters and numbers are more concrete than the descriptions, but the context provides no obvious thematic *link* between antecedent and consequent.

The findings of these experiments with thematic materials was that students reasoned far more in accordance with the logical competence model—choice of $\neg q$ increased and of both $\neg p$ and q decreased. The argument was then made that since the *form* of the abstract and the thematic conditionals was obviously the same, and the content made such a difference to performance, then logic (the theory of form) must be irrelevant to explaining how people reasoned. Hence the lack of attention to the vast literature on the variety of forms of conditional sentences. A literature which takes it as obvious that these conditionals are *not* of the same form.

After the early demonstrations of powerful effects of thematic material, there was a search for a characterisation of what thematic material 'works'. There were failures of replication of the transport problem and demonstrations that merely providing concrete material without thematic linkage was not helpful (Manktelow and Evans [19]). Nothing, after all, could be more 'concrete' than the vowels and consonants that appeared on the cards in the 'abstract' task. It is thematic linkage between them that is lacking in so-called abstract material. Griggs and Cox [14] showed that regulations provided particularly facilitating kinds of thematic linkage. Cheng and Holyoak [4] proposed that the thematic material that worked called up a repertoire of 'pragmatic reasoning schemas' citing examples such as permission, and obligation schemas. Claims were made that the only kind of thematic material which worked was 'social contract' rules (Cosmides [5]), and this for evolutionary reasons.

We will distinguish social contract thematic material as based on *deontic* conditionals (usually worded with *must*) from *indicative* rules which are descriptive. We will thereby mean to distinguish obligations from descriptive regularities rather than the particular grammatical moods that appear. It is quite common for indicative mood conditionals to be interpreted with the deontic force.

This social contract thesis was further refined by the claim that normative performance was only facilitated by a combination of social contract rule, plus a suitable 'social role perspective' (such as rule enforcer, or rule beneficiary) (Gigerenzer and Hug [12]). For example, Gigerenzer found that the rule "If the hiker stays overnight, he must bring fuel' with the subjects task being "to turn cards which must be turned to see if they obey the rule", produced relatively good performance when subjects were instructed to adopt a 'policing perspective' (imagining having the job of enforcing the regulation), and substantially worse when instructed to adopt what might be called an epistemic stance (seeking to decide which of two regularities pertained (perhaps the fuel was brought by guides rather than hikers)).

There have been counter claims that good performance can be achieved without resorting to deontic material and particular social perspectives. Sperber et al. [24] provide a fault finding scenario in which an engineer is seeking to find out whether a machine is printing cards correctly and this material produced good performance in at least some sub-conditions, though it is interesting that an apparently similar experiment by Griggs [13] earlier failed to find such facilitation. Sperber's experiment might be argued to have a rather leading hint about seeking a particular type of exceptional instance. However, Almor and Sloman [1] used thoroughly non-deontic material which might best be described as incorporating qualitative laws of physics and obtained good performance from their student subjects.

Early on, Wason, in several papers, investigated the relationship between insight and reasoning by using interviewing and thinking aloud protocols. He distinguished three levels of insight. Subjects conceived the task in terms of a search for : 1) just positive instances; 2) positive and falsifying instances; 3) only falsifying instances. He also distinguished two kinds of feedback: 1) feedback from hypothetical turnings —"suppose there is a 7 on the back, what would you then conclude about the rule?"; 2) actual feedback in which the subject turns the card and finds the 7. The most striking data from these studies is the observation of states of apparent inconsistency. The subject hypothesises (or discovers) an A on the back of 7, and notes that this would mean the rule was false of the card, but then declines to choose the card (or revise an earlier failure to choose it). In fact, as the evaluation and construction tasks have shown, *reasoning* about the cards does not seem to be the problem. Wason even reports that subjects can normatively justify card choices when those choices are presented to them (rather than elicited from them). One of Wason's strongest empirical claims from these studies of insight is that subjects who start with the p and q choices, *never* do reach a state of complete insight.

3 What has to be explained?

If these are the bare outlines of the observations, it is worth pausing to ask what needs to be explained. What are the desiderata of a cognitive theory of performance in this task? In particular, how should such a theory fit into the larger landscape of cognitive theories? Some might counsel that such questions are premature—"let's first understand the 4-card task before we start speculating about larger landscapes". But it is arguable that exactly the opposite strategy is best. If Wason's task only makes sense in the larger cognitive context, then we will never understand the task until we consider its embedding in that context. Certainly, the lessons of the observations are unlikely to have much interest for those outside the sub-sub-field unless they can be related to rather grander concerns. The fruit fly is interesting because a theory of genetics connects it to grand questions of nature and nurture. What are the relevant connections for Wason's task?

One obvious candidate is the issue of form and content in information processing—in particular human communication and reasoning. The analysis of the form of representations is virtually constitutive of understanding communication and reasoning. The ability to assign the same form to two token representations is a minimal requirement for any theory of communication or reasoning. Phenomena start by being described in contentful terms, and theory makes progress just as the analysis of form advances and encompasses explanations of observations.

Psychologists have sometimes been tempted to speculate that people can only reason with conditionals which they can remember from past experience. At one point, observing that Plymouth undergraduates perform differently with a transport problem than their London peers, Newstead speculates (perhaps half in jest) that this could be explained by supposing that they can retrieve reasoning about one transport destination from memory but not another. Such a 'memory' theory of conditional reasoning has redoubtable problems with understanding the processes of cognitive development and of transfer of reasoning, but that aside, note that such a theory still has to assign *form* and distinguish it from *content.* Memory representations must have form in virtue of which they are stored and retrieved, just like the representations of any other language.

But form is always relative. *If there is a vowel on one side, there is an even number on the other*; *If it's a mammal, it has hair*; and *if you are under 18 you mustn't drink alcohol* may share form at one level of analysis and diverge in form at another. In a natural language, almost any feature of the linguistic or non-linguistic context of utterance of a sentence may play a role in determining the form of the interpretation which it gets assigned and the processes that work over that form. So form will interact with content in reasoning, and the goal of theory is to understand how. The best that an observation that content affects behaviour can establish is that a deeper analysis of form is required. This *form* may or may not be linguistic *form.* It may be the form of the context; of the memory; of the task; of any factor controlling behaviour. But theory demands form, because only form has generality.

Semantic analysis of conditionals provides theories of sentential form. Unfortunately, the simplest grossest form assigned is that of the material conditional, and this for reasons of the history and pedagogy of logic. Cognitive development recapitulates *in reverse* the history of logical developement. Understanding truth functions comes last. It is an achievement of late stages of formal education, nowadays shunted into a particular specialisation which relatively few students take. Fortunately, the philosophy of science provides a sustained study of much commoner uses of conditionals understood by all—so called *law-like* conditionals, and in more recent times formal logical theory has to some extent caught up with the earlier stages of cognitive development in providing a range of analyses of modal counterfactual interpretations of law-like conditionals.

So the first thing we would like a theory of Wason's task to explain is how the various circumstances of the task and features of the subjects, control the assignment of forms to rules, tasks and contexts, and the part this assignement plays in determining reasoning and choice. We would like to connect theories of the forms of sentences to theories of reasoning with sentences.

One feature of this requirement is perhaps worth distinguishing as a requirement of its own, if only because it has so far been so thoroughly neglected. That is the contrast between what different subjects do in the same version of the task. Discussion has almost exclusively been about what circumstances increase the number of subjects showing normative behaviour. But in every version of the task, subjects exhibit a range of behaviour. Repeating the task on the same subject typically shows a tendency for the same behaviour to be repeated (see, for example, Gebauer and Laming [10]). It is a feature of cognitive theory at its current stage of development, that it tends, quite rightly perhaps for a new endeavour, to focus strongly on what is universal about subjects' behaviour, beneath surface variety. But if there is systematic difference between individuals in reasoning, then explaining this is both a desideratum of theory, and a tool of analysis. Comparing reasoning processes may be easier than providing absolute analyses.

One of the few sustained attempts to analyse the differences between subjects' reasoning is the early work of Wason and Johnson-Laird [29] in which they use in-depth interviewing of subjects about the reasons for their choices, and their responses to both hypothetical and actual consideration of the hidden sides of cards. An interesting claim made in this research was that subjects who initially choose the p and q cards only, never reach a state of 'insight' about the task during these socratic tutoring procedures. This is a case of a claim about individual differences between subjects. A satisfactory theory would, of course, have to connect some feature of these subjects' abilities and experience with features of the experiment to explain why the two different kinds of behaviour arise. So the second thing we would like is some understanding of individual differences in performance.

Wason's early investigations of insight also point to a third desideratum for theory. We would like to understand the processes involved in insight themselves—what might be termed the *phenomenology* of the task. Running subjects in this task generates 'aha!' experiences (as well as 'oh damn!' experiences). As subjects are exposed to either hypothetical or actual conflict between their reasoning and the cards, *some* of them have vivid experiences of insight or appreciation of error, and these are sometimes accompanied by abrupt shifts of reasoning and changes of explanation. A full theory of performance in the task would be able to explain the relationships between reasoning and these experiences. Completeness here is, of course, a tall order. But at least there must be room in a theory to explain these relations. They focus attention crucially on the relationship between competence and performance theories. If some subjects experience themselves as having made, and come to see through, what they themselves come to consider as errors in their reasoning, then it is a bold theory which denies that they earlier made an error. So thirdly, we would like a theory which linked reasoning and learning to experience of reasoning and learning.

One particular kind of individual difference that is perhaps worth special attention is differences in educational experience, both before and after learning Wason's task. Broadly, differences are observed between students who have studied different subjects, or studied them in different ways. But narrowly, and perhaps more interestingly, the process of learning to achieve normative competence in this task would seem to be an interesting cognitive process in itself. An understanding of this process might make a contribution to our more general educational understanding of increasing competence in reasoning tasks. And what impact does learning to do the task 'right' then have on what else students can do? There are those who would argue that learning to perform silly laboratory puzzles teaches precisely the performance of silly laboratory puzzles, and nothing more. A different view is that learning to be able to operate in the curiously abstract circumstances of laboratory experiments cut off from the rich cues of at least some circumstances of everyday reasoning, could play a part in a suitablly designed pedagogical program for the teaching of the 'core-skills' of reasoning. This is an old educational debate about transfer of formal knowledge. Any evidence for either answer would be of great applied interest, and would also feed into the theoretical debate about form, content, and transfer.

Finally, under desiderata for theory, we should say something about what has come to be known as rational analysis (see Anderson [2] and below). It is of course a vital psychological principle that theories should try to make sense of their subjects behaviour, if only because theories that make nonsense of it are likely to be false. The most insightful kind of explanation of non-normative behaviour explains 'error' in terms of conflicts between the circumstances of experiment and the circumstances to which subjects are 'naturally' adapted. An example of this type of explanation is the explanation of visual illusions in terms of perceptual mechanisms working hard in circumstances to which they are ill-adapted; cf. Marr [20, p. 294]. (There is, of course, always the possibility that the competence theory is just inappropriate.) Throughout what follows we will be engaged in rational analysis in this sense. We seek explanations of laboratory behaviour true to extra-laboratory interpretations of conditionals and associated instructions. Explanations which make the subjects' behaviour comprehensible, if not always 'correct'. We take seriously subjects' own categorisations of their reasoning as correct or fallacious.

In this we are no different from recent more specific interpretations of rational analysis (Anderson [2]; Oaksford and Chater [22]) which propose an application of Bayes' theory as the correct realisation of rational analysis. We shall see that under their analysis, subjects 'fallaciously' import assumptions about rarity of the truth of predicates in 'natural' circumstances into a laboratory task in which they are explicitly instructed that the rule is 'only about these four cards' and that frequencies are not as assumed in the natural environment. Different theories will offer different classifications and explanations of error. But there is no escaping the category of error, even more so since it is a category subjects freely use of themselves in explaining their own behaviour. Where we differ from rational analysis in its present incarnations is that we believe judgements of error indicate that there is no uniform notion of rationality at work here, and that these shifts of perspective should also be theoretically acccounted for.

4 Explanations

We group the various explanations of performance in the 4-card task under the headings: *matching bias*, *non-standard interpretation*, *familiarity*, *verification bias*, *matching bias*, *social contract theory*, *perspectival*, *Bayesian* and *task semantic* explanations. In this section, we discuss all but the last two explanations, to which we devote separate sections. As is so often the case, these explanations are not all mutually exclusive and can be classified in ways which bring out their similarities and differences. We will do this as we go along.

4.1 Matching Bias

Evans (see for example the review in Evans, Newstead and Byrne [7]) defines 'matching strategy' as the choice of cards which match the positive part of the content of a clause in a rule. So for the rule *If p then q*, p and q cards match: for the rule *If p then not q* still p and q cards match: and the same for *If not p then q*. Evans conceptualises the use of this strategy as a 'superficial' response to both rule and task which subjects adopt prior to processing the information to the level of a coherent interpretation of the whole sentence. As such, the strategy may be applied prior to, or alongside other processing strategies. It is taken to explain the modal response of turning the p and q cards in the abstract task. It must assume that something else is going on (perhaps superimposed on matching) when subjects adopt other responses. Thematic effects have to be explained in terms of contentful processes engaging other processes at deeper levels than matching.

Oaksford and Stenning [21] by investigating a full range of clause negations in both selection and evaluation tasks, showed that matching is not a particulary good explanation of performance with the full range of negated conditionals. They argue that a better summary of the data is in terms of the degree to which the material and instructions allow negative clauses to be processed as corresponding positive characterisations.

But perhaps the basic problem with matching is the difficulty of falsifying the theory, and whether the kind of truly superficial processing which people undoubtedly can engage in is really the interesting behaviour to investigate, granted that deeper processing can easily be induced to go on.

4.2 Interpretation and reasoning

When non-normative performance is observed in a psychological experiment, it is generally open to the experimenter to question the subjects' interpretation of the materials or task. Indeed, it is incumbent on the experimenter to ensure that the interpretation is as claimed for any subsequent theoretical deductions. There is a long history in the psychology of reasoning of explaining performance in terms of what we will loosely call non-standard interpretations, by which we will mean any interpretation signifcantly at variance with from the one assumed by the experimenter. Henle [15] is perhaps the most extreme proponent of this approach, claiming that virtually all divergence from normative reasoning is due to divergence of interpretation. Early in the 4-card literature, Wason [28] considered the possibility of 'biconditional' interpretation of the conditional, and Bracewell and Hidi [3] proposed that the 'one (side)' anaphor in the rule might be interpreted in a constant rather than a variable reading. A 'constant' reading of the anaphor results in an interpretation which can be paraphrased: 'if there's vowel on the visible side of the card, then there's an even number on the back.' Adopting this interpretation (along with a conditional rather than biconditional reading) would explain subjects' choosing just the p card.

More recently Gebauer and Laming [10] have used a modified method to argue that concrete anaphora and biconditional interpretations, both singly and in combination, are prevalent, persistently held, and consistently reasoned with. Gebauer and Laming present the four cards of the standard task six times to each subject, pausing to actually turn cards which the subject selects, and to consider their reaction to what is found on the back. Their results show few explicitly acknowledged changes of choice, and few selections which reflect implicit changes. Subjects choose the same cards from the sixth set as they do from the first. Gebauer and Laming argue that the vast majority of the choices accord with normative reasoning from one of the four combinations of interpretation achieved by permuting the conditional/biconditional with the constant/variable anaphora interpretations.

We would question how much persistence of choice means consistency of reasoning from an interpretation. The subject is given no feedback about the 'correctness' of their selections from the experimenter, and so might well feel there is a premium in consistency of selection. We know from the early 'insight' experiments that subjects are well able to persist in at least apparently inconsistent verbalised inferences. It is certainly true that Gebauer & Laming's subjects show that they are able to consistently categorise antecedents and consequents as true and false, but how much more we can infer about the consistency of their reasoning from this categorisation is a moot point.

We would also question the plausibility of the concrete anaphora interpretation of the English sentence without further context. Perhaps under the duress of reasoning about a number of complex possibilities required by the variable anaphora interpretation, some subjects to adopt an interpretation which simplifies their task. But out of context, the concrete anaphora interpretation appears to be an odd one. Here we are highlighting the possibility that interpretation and reasoning may be highly interactive processes.

The biconditional interpretation is a somewhat more complex issue. Geiss and Zwicky [11] have argued that the biconditional is the natural interpretation of many conditionals, especially deontic promises and threats. When I promise you "If you read this, I'll buy you lunch", I am at least dropping a heavy hint that no reading, no lunch. This hint appears to be generated on the roughly Gricean grounds of relevance. On the other hand, for non-deontic conditionals, biconditional interpretation while not impossible seems to stand in need of motivation. It might be that something like closed-world assumption reasoning might operate to generate this interpretation in experimental conditions. The very fact than no other rule is known might generate the inference that this is the only explanation. For example, "If the switch is up, the light is on" given without any further context, invites a world closed to other switches and therefore one in which the switch 'controls the light'—a biconditional interpretation. Providing a second rule "If Switch 2 is down, the light is on" might be sufficient to cancel this inference to biconditionality. The world has been augmented, the first switch no longer exercises total control over the light, and the relationship is now conditional but not biconditional. Such effects have been demonstrated in inference tasks by Byrne.

On its own, a purely interpretational hypothesis would suggest that a subject interpreting the rule biconditionally would turn all four cards. This is a rather rare event. Such a hypothesis hardly helps to explain the modal choice of just the p and q cards. Only when biconditionality is combined with a tendency to look for instances which make the rule true might it help to explain this modal behaviour.

4.3 Verifying and falsifying

This brings us to verification bias. This was Wason's initial explanation of his findings, which he took to be an application of Popper's claims in the philosophy of science. The first thing to be said is that there is a terminological issue about *verification*. If, as Wason believed, the only way to ensure that the rule is true is to seek falsifying instances, and verification means to seek instances which make the rule true, then verification is just what most subjects aren't doing (i.e. seeking falsifying instances), and it wouldn't be a bias if they were. In fact, on Wason's Popperian approach, verification and falsification are processes which differ only in their outcome, not what has to be sought. Wason clearly means by verification bias, a tendency to seek instances which comply with the rule—we might rename this *compliance* bias, but the term 'verification bias' is so well embedded in the literature that it is perhaps better to note the conflict with normal usage. This might be a quibble if there were not serious questions about how subjects interpret the task instructions, an issue to which we return below.

If subjects are seeking compliant cards, which cards are those? Clearly p/q cards are compliant. Clearly $p/\neg q$ cards are not compliant. Truth table elicitations from subjects might be interpreted as meaning that the majority of subjects regard both $\neg p/q$ and $\neg p/\neg q$ cards as neither compliant nor non-compliant but irrelevant. If we accept this interpretation, verification bias could explain the modal response. Subjects turn the cards, and only the cards, that could be compliant with the rule. And they turn them to see whether they are so.

This last claim might already provide a challenge to some interpretations of verification bias. If subjects turn, for example, p in order to see whether it has a q, but also to see whether it has a $\neg q$, then they must be construed as (at least partly) seeking non-compliance and therefore falsification. It is true that when *asked* why they chose p, they typically mention only that finding a q would confirm the rule, but it is not uncommon also to mention, especially under mild prompting, that finding a $\neg q$ would falsify. Omitting to mention this until prompting might well be taken to be explained on Gricean grounds of quantity. Subjects tend to be withering if asked what would follow from discovery of a $\neg q$. Certainly, it is most unusual (although it does occur) to find any subject who fails to explain the relevance of discovering $\neg q$ on the back of p when asked about it. This is in marked contrast with explanations about the $\neg q$ card. Further systematic investigation of subjects' explanations for turning p are needed.

So if verification bias is to be counted an explanation of the modal card choice, it needs the assumption that $\neg p/q$ and $\neg p/\neg q$ are irrelevant rather than compliant. Furthermore, it is questionable whether subjects fail to see the relevance of falsification in the case of card p, though they do appear to fail to see it for card $\neg q$. Finally, before leaving verification, we should mention that it shares several features with Bayesian explanations. The verification bias theory attempts to explain choice in terms of compliance: the Bayesian theory attempts to explain why seeking compliant cards is a good strategy for gaining information under various conditions. But more of Bayes below.

4.4 Social contracts and cheating detectors

So far, we have not considered how the various explanations explain thematic effects, save perhaps for our oblique reference to the idea that some deontic conditionals tend to be interpreted biconditionally. However, this particular link (that deontic conditionals tend toward biconditional interpretation) most certainly won't explain the main observations of reasoning with deontic conditionals. These are just the rules where card selection is most normative. Social contract explanations focus almost entirely on thematic effects. Cosmides [5] original claim was that human beings, during their social evolution, developed 'cheating detector' algorithms which functioned to allow them to police social contract regulations, and that it is *only* when these algorithms are brought into play that people can make the required inferences in the 4-card task. Cheating detectors are the only mechanism with which undergraduate students (prior to logical instruction perhaps) can solve the task.

This is an extraordinarily strong claim, and a rather curious application of evolutionary theory. We have absolutely no argument with the general importance of evolution in understanding human psychology. Even on a much shorter timescale, there are excellent historical arguments that the development of our understanding of law-like scientific conditionals developed historically from our understanding of deontic conditionals, in the sense that orginally natural laws were taken to be *ordained* in the same sense as legal laws.

We might also accept that language probably developed with an emphasis on social understanding and control rather than physical or causal understanding. The indicative may be the 'neutral' mood as evidenced by various linguistic criteria, but it is most unlikely to have been the original mood. But the important question is what bearing this has on modern undergraduates' performance in the four-card task. Granted that language has moved on, wouldn't it be extraordinary if intelligent undergraduates were incapable of say reasoning about evidence for causal conditionals in simple natural thematic contexts? Why should material which invites scenarios which embody social contracts and can be solved by 'cheating detection' be the only ones for which subjects can solve this task in the normative manner? In fact, we have seen that they are not. Several non-deontic contexts have been shown to facilitate normative reasoning, e.g. Sperber et al.[24] and Almor and Sloman [1]. Nevertheless, there are numerous demonstrations that providing simple thematic material of say a causal nature is not sufficient to bring out normative reasoning. There is certainly something to be explained about the role of the deontic/indicative moods in these observations. Before leaving this question, we note that whatever the mechanism of the 'deontic effect' in the selection task, it does not operate in evaluation or construction tasks. If some 'cheating detector-widget' were our only implementation of conditional reasoning, we would need to seek some other explanation of how we perform these other conditional reasoning tasks?

4.5 Perspective

Finally, we consider what we will call 'perspectival explanations' by which we group together explanations in terms of perceptual, or 'information packaging' explanations. One way of posing the fundamental puzzle of the 4-card task is to observe that the A-card, if it happens to have a 7 on the back, is the very same card as the 7-card if it turns out to have an A on the back. This very same card, viewed from opposite sides, calls forth different conceptualisations. Viewed from the letter side, it is easily chosen as relevant to turn: viewed from the number side it is rarely chosen. This is what we call a difference of perspective. Perspective appears to have some inertia—Wason and Johnson-Laird's [29] subjects who had already turned over the 7 and found an A still did not necessarily 'see' it as the same type of card as the A which they had turned over and seen to have a 7 on.

Perspective is a perceptual term, but in this context it is closely aligned with what linguists call, among other things, 'information packaging' (see e.g. Vallduvi and Engdahl [6]). Describing something as an 'A which has a seven' may call up a different representation than describing it as a 'seven with an A'. They may be propositionally identical but they are informationally distinct descriptions. For a related example, consider the fact that about 20% of naive undergraduate subjects deny that 'Some A are B' entails that 'Some B are A' (Stenning, Cox and Oberlander [26]).

Negation is a powerful packaging device. Describing a letter as 'not a vowel' induces a different perspective than describing it as a 'consonant'. Wason's [27] early work on 'contexts of plausible denial' illustrated this point. The four-card task is set up to allow translation of negations into positive forms by its background rule: "All the cards have a letter on one side and a number on the other". This is, not to say, that subjects immediately exploit this information. The perspectival effects of negation are closely allied to the Bayesian explanations of performance. The ravens Paradox turns on the propositional identity but informational distinctness of *All ravens are black* and *All non-black things are non-ravens* (see below).

There are also effects of temporal information packaging in the 4-card task, induced by the division of thematic material between antecedent and consequent. A causal conditional is most neutrally stated with cause in the antecedent and the effect in the consequent. This might be labelled the perspective of control or prediction. “If the switch is up the light is on”. But if we want to invoke a diagnostic perspective, we say “if the light is on, the switch is up”. The switch controls the light, but the light may tell us about the switch.

This perspective difference interacts with the task set up. At the outset, the cards are visible one side and invisible the other. So there is a temporal asymmetry induced by the idea of turning, which maps on to an epistemic asymmetry in terms of what is known/unknown, and onto an informational asymmetry in a grammatically structured rule. In the case of the A-visible card, all these asymmetries so-to-speak line up. The cause is known and is positively described in the antecedent which invites an inference which predicts an outcome of turning. In the case of the 7-visible card, everything is ‘inside out’: the causal state is what is unknown, and its known effect value is visible but negatively described in the rule relative to the card.

It is notable that the scenarios which get good performance with indicative rules uniformly invoke diagnostic perspectives. Both Sperber et al [24] and Almor and Sloman [1] have ‘quality control inspectors’ looking for compliance. Cheating detection is a kind of diagnostic perspective—one observes the behaviour which has taken place, and asks whether or not it conforms to a law.

Interestingly however, even with deontic rules, when a discovery perspective is induced (which of two rules explains behaviour), performance becomes worse (Gigerenzer and Hug’s mountain hut firewood experiment). Perspectival explanations are not best thought of as entirely distinct from the other kinds of explanation. There are close links between interpretational, Bayesian, social contract, matching bias, and verification explanations on the one hand and perspective explanations on the other. Explanations in terms of ‘relevance theory’ (Sperber and Wilson [25]) are perhaps best seen as perspectival explanations.

One might object to perspectival explanations in general that they are non-explanatory simply because any theory can be cast in terms of perspectives. There may be some truth in this complaint about specific cases, but we include them as a general kind of explanation here because of their crucial role in connecting performance to phenomenology. Subjects’ experiences of insight are very often described in terms of ‘seeing’ a card differently, and at least sometimes, experiences of insight are accompanied by changes in performance.

5 A Bayesian explanation

The point of departure of the Bayesian explanation, due to Oaksford and Chater [22] is that the 4-card task is first and foremost a problem about decision, not about logical reasoning. This makes good sense as a modelling strategy, for we have seen that subjects evaluations of the relevance of a card are to a large extent determined by their previous selections, even when this entails a conflict between the selection and logical reasoning. What then determines the selection process? In the Bayesian model, what matters is a subject’s subjective probability of the hypothesis that the conditional is true, given his prior information. It makes sense to talk of probability in the 4-card task if one assumes that subjects will misunderstand the experimenter's instructions by taking the four cards to be a sample from a larger population, whereas the intended interpretation of the instructions is that the rule pertains to the four cards only.

The essential consideration is then that selecting a card may be viewed as the selection of a possible experiment, testing the hypothesis. Now as in, say, a medical situation, we may compare experiments, i.e. card selections, in terms of their potential relevance to the truth of the hypothesis. More formally, we may compute the information about the hypothesis yielded by an outcome, and then average over the possible outcomes. It then seems sensible to choose the experiment with the highest expected information gain. In a nutshell, this is Anderson's [2] procedure of "optimal data selection", which is taken by him to underlie much of cognition. It is also known by the catchphrase 'rational analysis'. In a rational analysis of a particular cognitive activity, one tries to show that an organism's behaviour is optimally adapted to the environment, even though it may not conform to whatever canons of logicality apply. The general methodological strategy behind rational analysis is model fitting, i.e. proposing a statistical model involving a sufficient number of parameters, so that upon estimation of the parameters the model fits a collection of data points, namely the organism's behaviour. The function of the parameters is to succinctly characterise the organism's environment. Optimality then consists in maximising a number of standard measures, such as expected information gain, or expected utility, whose relevance to the organism are taken for granted. If one has thus succeeded in fitting a model to an organism's behaviour in a particular cognitive domain, one says that behaviour in this domain has been given a rational analysis. We shall come back to the normative status of this type of analysis below.

In any case, using several assumptions which allow one to estimate the probabilities involved, the computation of expected information gain yields the following rank order of cards to be selected

$$p \; > \; q \; > \; \neg q \; > \; \neg p.$$

This then is the proposed explanation of why the q card is chosen much more frequently than the $\neg q$ card. The reader might object that this explains rather too much, since in at least some concrete versions of the task, the rank order is

$$p \; > \; \neg q \; > \; q \; > \; \neg p.$$

This outcome is handled by adding utilities to the model; roughly (details will be given below), the abstract task is characterised by the fact that we are more or less disinterested in the outcome, so that the utilities are the same, whereas the concrete task is characterised by an uneven distribution of utilities.

We will now discuss the model in greater formal detail. Interestingly, it is adapted from what has been described as the solution of the ravens paradox, by Mackie. The ravens paradox is that observation of a nonblack nonraven confirms the statement that all ravens are black. The solution proposed by

Mackie is that one should compare *two* hypotheses: H_0 says that the properties 'raven' and 'black' are independent, whereas H_1 is 'all ravens are black', hence complete dependence. In probabilistic terms, we then obtain the following model. Suppose we believe the proportion of ravens is x, and that of black objects is y; independence then says that the probability of black ravens is xy; similarly the probability of nonblack nonravens is $(1-x)(1-y)$. In the case of complete dependence the situation is slightly diferent; since the probability of there being a nonblack raven is 0, we get that the probability of a nonblack nonraven is $(1-y)$. Now suppose we observe a golden oriole. Assuming for simplicity that the a priori probabilities for H_0 and for H_1 are equal, we obtain (representing this observation of a nonblack nonraven by 'O')

$$\frac{P(H_1|O)}{P(H_0|O)} \;=\; \frac{P(O|H_1)}{P(O|H_0)} \;=\; \frac{1-y}{(1-x)(1-y)} \;=\; \frac{1}{1-x}.$$

Mackie nows introduces the 'rarity assumption' which says that the relative frequency of positive predicates in the world is low. In our case, this implies that $x \sim 0$, so that $\frac{P(H_1|O)}{P(H_0|O)} \sim 1$, which entails that $P(H_1|O) \sim P(H_1)$. Of course, under the same assumptions, if BR denotes the observation of a black raven, $\frac{P(H_1|BR)}{P(H_0|BR)}$ will be huge; and if nonBR denotes the observation of a nonblack raven, $\frac{P(H_1|\mathrm{non}BR)}{P(H_0|\mathrm{non}BR)}$ will equal 0.

We may recast the preceding argument in terms of the information gained by, on the one hand, observing ravens to see whether they are black, and inspecting nonravens on the other hand. Let X be an experiment with two outcomes, X_0 and X_1. Then the expected information gain upon performing X, $E_X(I)$, is given by

$$E_X(I) \;=\; \sum_{i,j=0,1} P(H_i, X_j) \log_2 \frac{P(H_i|X_j)}{P(H_i)}.$$

Suppose first that X denotes the experiment which tests whether *nonravens* are black. If we observe that the nonraven is nonblack, outcome X_0, we have already seen that $\frac{P(H_i|X_0)}{P(H_i)} \sim 1$. If the outcome is X_1, the nonraven is black, then one easily computes that $\frac{P(H_0|X_1)}{P(H_0)} = \frac{y(1-x)}{y-\frac{1}{2}x(y+1)}$ and $\frac{P(H_1|X_1)}{P(H_1)} = \frac{y-x)}{y-\frac{1}{2}x(y+1)}$, which, given the assumption on x are both of the order of 1. It follows that $E_X(I) \sim 0$. On the other hand, if X denotes the experiment which tests whether *ravens* are black, we see that $E_X(I)$ equals $-\frac{1}{2}(xy \log_2 xy + x \log_2 x + x(1-y) \log_2 x(1-y))$, which will be larger. In sum, therefore, it makes more sense to test ravens for blackness than to test nonravens.

Let us now apply this line of reasoning to the 4-card task, pertaining to the implication $p \rightarrow q$. It is fundamental to Oaksford and Chater's [22] reconstruction that they assume that a subject interprets the conditional as pertaining to a population from which the four cards shown are only a sample. Of course, this was not the way the task was intended, but by thus misinterpreting the task, the subjects naturally brings in probabilities and rival statistical hypotheses. Selecting a card and turning it over can be viewed as performing an experiment. As in the case of 'all ravens are black', the experiment is brought to bear on two rival hypotheses, H_0 stating that p and q are independent, H_1 asserting that p is included in q. Accordingly, each card, determined by its visible side which is $p, \neg p, q$ or $\neg q$ also determines an experiment, and hence the expected information gain associated to that experiment, denoted by $E_p(I)$ etc. The rank order of the various $E_X(I)$ now depend on the probabilities $P(p),\ P(q)$[1], as follows:

1. if $P(p),\ P(q)$ are small (≤ 0.15), $E_p(I)\ > E_q(I)\ >\ E_{\neg q}(I)\ >\ E_{\neg p}(I)$;
2. if $P(q)$ is small, but $P(p)$ is large, the ordering obtained is $E_p(I)\ >\ E_{\neg q}(I)\ >\ E_q(I)\ >\ E_{\neg p}(I)$.

Oaksford and Chater argue that in the abstract case, the assumption of 1 is satisfied, and conclude from this that subjects do well in preferring to turn the q card over turning the $\neg q$ card. Many questions remain, of course, among which the following stand out

1. subjects typically choose a *set* of cards, not just a single card; what is the rank order of expected information gain associated with these more complex experiments?
2. what are the predictions of the model when the rule is varied by introducing negations in antecedent and/or consequent?
3. what is the difference between the abstract and the thematic tasks?

Question 1 is motivated by the consideration that subjects, instead of the logically correct choice $\{p, \neg q\}$, may choose sets such as $\{p\}$, $\{p, q\}$ or $\{p, q, \neg q\}$. To take account of Wason's results, the model would have to explain that each of the last three experiments has higher expected information gain than the first experiment. The second question is interesting because of its interaction with the rarity assumption. Take the case of a negative antecedent, for example the rule 'if there is not a vowel on one side, then there is an even number on the other side' ($\neg p \to q$). The rank order of responses here is $\neg p\ >\ q\ >\ \neg q\ >\ p$. In order to explain this rank order along the lines sketched above one would need a rarity assumption saying that $P(\neg p),\ P(q)$ are small. Now it seems clear that $P(\neg p),\ P(p)$ cannot be simultaneously small. Oaksford and Chater [22] offer two solutions here. The first derives from Oaksford and Stenning [21] and consists in interpreting $\neg p$ as an antonym of p, denoted $\sim p$, for which we may have $P(\sim p) + P(p) < 1$; in particular, Oaksford and Chater assume that $P(\sim p)$ is always ≤ 0.5. This move finds support in linguistics, but it does not solve all problems. The model imposes several boundary conditions on the probabilities; for instance if H_0 is independence of p and q, and H_1 inclusion of p within q, then one must have $P(q) \geq P(p)P(H_1)$. This is so, since (a) we may assume p to be independent of $\{H_0, H_1\}$ (otherwise observation of $p, \neg p$ cards could provide information about the true hypothesis) and (b) $P(q|H_1) \geq P(p)|H_1)$ by definition of H_1. By the same token, however, the model set up to explain subjects behaviour with respect to the rule $\neg p \to q$ forces the inequality $P(q) \geq P\neg(p)P(H_1')$, where H_1' says that $\neg p$ is contained

[1] Strictly speaking one also has dependence on $P(H_0)$ but the rank order is by and large independent of this value.

in q. This bounday condition is easily violated when p, q are rare. Oaksford and Chater propose that, faced with this inconsistency, subjects revise their estimates for P(p) upward, and they adduce the fact that subjects have more difficulty comprehending the conditional $\neg p \rightarrow q$ (as measured by reaction times) as support for this proposal.

Lastly, answering question 3 requires a considerable extension of the model. We have seen above that in a thematic, in particular deontic, task the range of reponses can be greater, depending on the *perspective* that the subject adopts. To recapitulate, deontic rules come in two varieties, *permissions*, 'if condition (p), then may action (q)' and *obligations*, 'if action (p), then condition (q)'. Deontic rules may be viewed from two perspectives, that of an *actor* and that of an *enforcer*. We illustrate this for the permission rule. The actor tries to perform the action by satisfying the condition; from his perspective, the rule is violated if the condition is satisfied, but the action does not take place (p and $\neg q$). The enforcer tries to see to it that *only* people who satisfy the condition perform the action; for him the rule is violated in case $\neg p$ and q.

Oaksford and Chater model this by assigning utilities depending upon perspective, in such a way that from the actor's perspective, the maximum expected utility of $p \wedge \neg q$ is highest, whereas for the enforcer maximum expected utility is assigned to $\neg p \wedge q$. In order for this to work, the rarity assumption must be dropped. The abstract task arises as a kind of limiting case of the deontic task, where the perspective is that of a dispassionate enquirer who assigns equal utilities to all outcomes.

5.1 Methodological animadversions

The virtue of Oaksford and Chater's approach is that it is an ambitious attempt to explain all phenomena pertaining to the selection task within a single model. As such, it is without equal. However, even the cursory review of Oaksford and Chater's model given above will have made clear to the reader that the model involves many free parameters and assumptions. Many more assumptions can be found strewn across the footnotes or in parenthetical remarks in the main text. The aim was to fit a model to the data, but this is always possible if the model contains enough free parameters. In this case the situation even appears to be slightly worse; we have seen, while discussing negated antecedents, that the authors felt obliged to change parameters values in mid-argument. Surely not all such moves can be justified by pointing to changes in the environment, as a rational analysis requires.

In this respect it is of interest to discuss Oaksford and Chater's [23] reaction to an experiment of Pollard and Evans (for a discussion, see Evans and Over [8]), which at least at first sight appears to be a test of this particular Bayesian model. Pollard and Evans manipulated the conditional probability $P(q|p)$ (which they equate with the probability of the conditional $p \rightarrow q$) with a view to demonstrating that if the conditional is usually false, i.e. if $P(q|p)$ is low, then subjects are more likely to choose the $p, \neg q$ cards. The manipulation consisted in showing subjects two sets of cards. One set (for the ususally true conditional) was composed of seven p, q cards, one $p, \neg q$ card, seven $\neg p, q$ cards and seven $\neg p, \neg q$ cards. The second pack had one p, q card and seven $p, \neg q$ cards, but was otherwise the same. Participants are shown one face of the card, are asked to predict what is on the other side, and then turn the card over. It indeed turned out to be the case that in the usually false condition subjects are likely to choose $p, \neg q$ cards. This was explained by memory cueing: if the conditional is usually false, the subject will have seen more counterexamples. As such this is not incompatible with a Bayesian account, but it seems to be incompatible with an analysis in terms of expected information gain. This is so, roughly, because a usually false conditional will have low a priori probability, which will move toward 0.5 upon confirmation, which for the entropic measure of information used counts as an increase in uncertainty. Consequently, the expected information gain for turning the $\neg q$ card is very much smaller in this case than when the a priori probability of the conditional is high. The upshot is, that Oaksford and Chater would have to predict that more $p, \neg q$ cards are chosen in the usually true condition, which, as we have seen, is not true. Their way out is, first, to argue that a Bayesian should not be dismayed by a single falsification of his theory, and second, to observe that in the usually true condition the rarity assumption is violated; since the subjects explicitly learn, in the training phase, only the conditional probability $P(q|p)$ and not the actual values of $P(p)$ and $P(q)$, they might adopt default rarity values for $P(p)$ and $P(q)$, thus cancelling the prediction that the usually true conditional would lead to a high proportion of $p, \neg q$ selections. This is a clever but suspect move, since it would seem that subjects cannot fail to estimate the true values of $P(p)$ and $P(q)$ from the data.

5.2 Rationality and logic

Of course, the main issue of interest to us is the relation of Bayesian analyses to logic. To sum up, a 'rational analysis' along the lines of Oaksford and Chater claims to show that the rank order $p > q > \neg q > \neg p$ is actually better adapted to the real world than the logician's choice. A few remarks about the concept of rationality are in order here. It has been customary to describe the observed result in the abstract version of the selection task (only 4% chooses the correct $p, \neg q$ cards) as pointing to human 'irrationality'. Even from a logicist's viewpoint it seems better to make a distinction between 'bounded rationality', failure to consider a particular inference pattern (here the relevance of the $\neg q$ card), and 'irrationality' in the sense of selecting a card (q) which is not relevant to the truth of $p \rightarrow q$. This distinction could make sense of Wason's observation that subjects who choose p, q originally hardly ever get complete insight upon tutoring, whereas this is not so for subjects whose original choice is p only. It might seem that bounded rationality and irrationality call for different explanations.

The Bayesian approach takes for granted that it is rational to maximise expected information gain and expected utility, apparently more rational than applying modus tollens. Even assuming that this so, then, as Laming [18] rightly points out, there is something curious in the way Oaksford and Chater use Bayesian criteria of rationality: if turning the p card has highest expected information gain, then subjects should *always* perform this experiment, not just in a large percentage of cases. Similarly, the Bayesian injunction to maximise expected utility is a rule which should always be followed, not most of the time, so that in the thematic case all subjects would have to choose the $p, \neg q$ cards. The upshot is that the rational analysis shows only that a certain percentage of subjects is adaptively rational, not that each and every human is. This makes the criteria for the adequacy of a rational analysis even more vague.

A deeper problem is that the Bayesian concept of rationality is as it were parametrised. A trivial instance of this is that maximising expected utility depends upon the assignment of utilities. Assigning nonuniform utilities merely points to situations where falsification is all-important and seems to be a description of the situation rather than an explanation. Without further inquiry into why assignments of utilities differ in the abstract and concrete task, the model is not very informative. As we will see below, merely changing the intended utilities in the abstract task does not change behaviour. Less trivial is that one may be easily bullied into submission by the expression 'maximal expected information gain'. Who would be against maximising information?

The trouble is, of course, that the formal notion of information is just a mathematical construct, perhaps adequate to the task, perhaps not. The crux of the Bayesian argument is that, whereas information gain $I_g = I(H|D) - I(H)$ upon falsification is enormous ($\frac{1}{2}$ with the chosen prior), expected information gain is still small because under the rarity assumption the probability of a falsifying event is small. But this is just a consequence of the preliminary decision to represent all information on the same numerical scale. That decision is perhaps reasonable if the goal is to assign a 'degree of confirmation' to an hypothesis, but it is much less reasonable if the goal is to establish the truth or falsity of this hypothesis. In that case, if $I(H|D) = 0$, we know all we wanted to, and one should give I_g a point of discontinuity by putting $I_g = \infty$, which choice, of course, destroys the expected information gain argument.[2]

In sum, we are back to square one. The Bayesian analysis has some plausibility (modulo the reservations expressed above) if the subjects construe the task as that of obtaining evidence on an hypothesis, in the sense of a posteriori probability. That is not how the task was intended, however. It is therefore not a matter of competing notions of rationality, let alone of vilifying falsificationism in favour of Bayesianism; the issue is rather how subjects understand the task. It is thus more profitable to obtain detailed information on a subject's semantics for expressions such as 'determine whether ... is true or false' and only then to consider whether the eubject's behaviour is rational given his semantics. This point will be amplified below.

5.3 Probability is a poor language!

The most damning objection to the Bayesian explanation is that adherence to probability theory paradoxically forces a too narrowly 'logical' account of the conditional. The conditional is modelled either by inclusion, or by inclusion

[2] Observe that the demonstration that his information function is the only one satisfying some plausible criteria, assumes continuity. See for example Khinchin [17], theorem 1.

modulo a small set of exceptions, where in the last case we need to refer to a probability measure. A priori it is rather doubtful whether the wealth of conditional meanings that logical and linguistic analyses have uncovered can be expressed in this parsimonious language. More importantly, there exists experimental evidence which shows that such a unitary account of the conditional fails to do justice to the facts. The evidence has to do with subjects' behaviour with respect to logically equivalent forms of the conditional. We consider first van Duyne's 1974 experiments. He compared four different formulations of a conditional statement in a both an abstract and a thematic task. In the latter case, the rules given were

1. implicative 'If a student studies philosophy he is at Cambridge',
2. universal 'Every student who studies physics is at Oxford',
3. disjunctive 'A student doesn't study French, or he is at London',
4. conjunctive 'It isn't the case that a student studies psychology and isn't at Glasgow',

and similarly for the abstract task. His idea was to compare the gains in insight for the four sentence types, when moving from abstract to thematic material. A priori, the predictions were as follows:

1. overall, there would be a significant difference between abstract and thematic materials;
2. in the abstract condition, the disjunctive formulation 3 would yield a higher percentage of correct selections since its unfamiliar form might draw attention to its logical properties;
3. in the realistic condition formulations 1 and 2 were supposed to yield more insight than 3 and 4, because the unfamiliar form of the latter may now override the thematic materials effect.

The first prediction was confirmed. The second prediction was not borne out, and subjects performed as badly in 3 as in the other formulations, in the sense that the percentage of correct answers is the same.[3] The third hypothesis was strongly confirmed however: the higher percentage of correct answers in the thematic condition was entirely due to gain of insight with the universal and implicative sentence types.

This result is highly relevant to our concerns. It shows that one cannot naively take the logical form of a sentence and use it in one's model *as if this were also the meaning assigned to the sentence by the subject*. For if this were so, all sentence types would perform equally well in the thematic task. One would therefore have to argue that subjects distort the meaning of $\neg$ and $\vee$ so that 1 is no longer equivalent to 3; but then there is no guarantee that subjects' meaning of 1 or 2 is precisely the logical meaning.

At long last, we are now able to formulate what is wrong with the Bayesian model *and* to explain the title of the paper. The fundamental mistake of the Bayesian model is that it is by and large insensitive to meaning. The moral is

[3] Interestingly, the 'matching response' p, q occurs much less frequently in formulation 3.

that psychology cannot proceed without paying heed to semantics. This has immediate consequences for the design of experiments. In the customary form of experiments involving a 4 card task, subjects receive an instruction what to do, for example 'select only those cards you definitely need to turn over to find out whether they violate the rule or not' (van Duyne), or 'your task is to say which of the cards you need to turn over to find out whether the rule is true or false' (Wason and Johnson-Laird), and then the subjects perform the task. Generally, the only aspect of subjects' behaviour which is then scored is the set of cards they select.[4] This 'bare' design is not likely to give us the necessary information about subjects' understanding of the key expressions involved. A different design is therefore called for; this will be explained in greater detail in section 7.

We now turn to the semantic characteristics of the 4-card task. In the next section, we will review empirical work describing what might be called the psychosemantics of the conditional, and in the following section we propose a new analyis of how the semantics of various kinds of conditionals interacts with understandings of different instructions.

6 How subjects may understand the rule

As we saw in the section on interpretation, the designers for the 4-card task were in some measure aware of the possibility of misinterpretation of the conditional, and some included instructions to prevent a bi-conditional interpretation. However, this appears to be *le moindre de nos soucis*, since it is the notion of conditional itself which is at issue, so that the instructions merely have the effect of preventing the *bi* of something unknown.

A logician aims at a theory *form* of the conditional which treats it as a logical particle—an expression whose meaning is independent of that of the propositions it connects. But the theory must encompass the contextual variations of the way the conditional is understood in natural language, a point already made by Wason and Johnson-Laird in [30]: the conditional "is not a creature of constant hue, but chameleon like, takes on the color of its surroundings: *its meaning is determined to some extent by the very propositions it connects*." We give a brief synopsis of the results of Fillenbaum's [9] experiments on the use of conditionals, corroborating this thesis.

A rough typology of the use of conditionals is given by

1. temporal-causal: 'if he bungles that job, he will be fired'
2. conditional promise: 'if you wash my car, I'll buy you an ice cream'
3. conditional threats: 'if you come any closer, I'll shoot'
4. contingent-universal: 'if the yellow light is on, the cab is for hire'.

Where does the conditional 'if there is an A on one side, then there is a 4 on the other side' fit in this typology? The only possibility would be under the contingent-universal heading. It turns out, however, that humans overwhelmingly (92%) find conditionals where antecedent and consequent are topically

[4] Van Duyne also scores judgements of similarity of the various guises of the conditional.

unrelated, as in this case, strange. This shows that the ordinary understanding of a conditional is context sensitive, and may be taken as an indication that reasoning performance is likely to be erratic, as subjects desperately try to attach a meaning to the conditional. Clearly, however, this cannot be the whole story, since the thematic materials effect also fails to appear in some *bona fide* contingent-universal conditionals.

Even given that antecedent and consequent of the conditional are topically related, it appears that the logical properties of the conditional are to some extent affected by the meaning of the constituent expressions. Fillenbaum tested this by eliciting paraphrases of the various types of conditionals, where the use of 'if (- then)' in the paraphrase was forbidden. It turns out that there is a consistent difference in paraphrases offered between the various types of conditionals. Promises are likely to be reformulated by means of a conjunction: 'wash my car and I'll get you an ice cream'. Disjunctions are hardly ever used here: 'don't wash my car or I'll buy you an ice cream' sounds distinctly odd. In the case of threats, on the contrary, paraphrases involving disjunctions abound: 'don't come any closer, or I'll shoot'. Now while conjunction can occur in paraphrases of positively formulated threats: 'come any closer and I'll shoot', this is much less acceptable for threats with negative antecedent: 'don't pay me immediately and I'll sue you' is a less likely paraphrase of 'if you don't pay me immediately, I'll sue you' then 'pay me immediately or I'll sue you'. Interestingly, temporal-causal and contingent-universal conditionals were almost never reformulated by means of disjunction or conjunction. The most common type of paraphrase was that in terms of a simple sentence; for example 'if the bed is soft then the man will sleep well' tends to become 'the man will sleep well on the soft bed'. This may point to systematic differences in meaning, and in any case provides no ground for simply equating 'if - then' with material implication or even with an implication which allows exceptions.

While the meanings of 'if - then' are already diverse, the problem is compounded by the fact that subjects tend to have very different views of what it means for a conditional to be false. Specifically, Fillenbaum asked subjects, what a speaker could mean by replying 'No, that's not so', to a given target conditional. The findings were most interesting. The denial of a contingent-universal conditional ifp, q ('if the yellow light is on, the cab is for hire') was taken to be *if p*, *then not q* in about 30% of the cases, and as *if p*, *then maybe q*, *maybe not q* in 60% of the cases. Reformulating this conditional as a universal statement ('all cabs with yellow light on are for hire') made a dramatic difference: 98% of subjects now gave as denial an expression of the form *Some p are q*, *some p are not q*. While this already shows that simple extensional treatments of meaning are grossly inadequate as underpinning for psychological theorising, one would also think that subjects understanding of denial would influence their understanding of the experimenter's instructions.

7 Task semantic explanations

We discussed above explanations in terms of alternative interpretations of the rule. What we here call 'task semantic' explanations can be thought of in terms of alternative interpretations of the instructions and how they apply to the materials presented to subjects. They might also be thought of in terms of differences in perspective on the semantic relations between rule and card or card and rule. The subject is asked to make decisions about turning cards on the basis of their relation to the rule, or to make judgements about semantic properties of the rule on the basis of the cards. But the possible semantic relations between rule and card are varied, asymetrical and complex, and are notoriously subject to contextual influences on interpretation.

The major contrast we focus on here is between the different semantic relations between card or cards and rule with descriptive and deontic rules. To take the clearest case first, with a rule expressing a (legal) law, when the law holds, a card can describe a case that obeys the law or one that violates it. Given that a law exists, nothing that we can discover on any card or set of cards could possibly tell us anything whatsoever about whether the law is, in fact, in force (notice particularly that the nearest property to *truth* which these kind of laws have is *existence*). The drinking age in Grigg's home state may or may not be 18, and the rule may or may not be intended to be interpreted here or there (or anywhere else), but whatever appears on the cards we will be no wiser. In fact, a law might *never* be obeyed but might still remain a law. Note also that the asymmetrical relation between law and case is typically described using distinct vocabulary: cases obey or disobey laws; laws either permit or prohibit cases.

It is interesting in this context that when Gigerenzer & Hug's instructions turn the task with deontic material into an 'epistemic' one of deciding which of two laws exists in a context, then performance with deontic rules declines markedly. Precisely when the semantics of the task is instructed to be descriptive, the subjects' problems resurface.

The same is true, though less transparently so, for rules expressing natural laws. Newton's first law can be stated in the form "If a body is in uniform motion unaffected by outside forces, it continues in motion in a straight line at uniform velocity". We can think of the cards as describing experimental observations, and these observations may either accord or conflict with the law, but if they conflict, we resolve that conflict by appeal to some violation of a background condition—there is a hidden force, etc. etc. So no card can show that the law is false. An accumulation of such cases might throw doubt on the generality of the law, but even when an alternative law puts clear bounds on generality, we are still likely to describe the old law as true in restricted circumstances (at suitably low velocities etc.).

In contrast, a rule which is taken to be descriptive (as opposed to law-like) rule, is shown to be false on finding a single counterexample. A rule may be true of a case, or false of that case. But here it is not nearly so clear what the inverse of these relations corresponds to, nor exactly what the inverse relation applies to. No card can 'make a rule true', though a set of cards which constitutes an entire domain may make a rule true. A single card may 'fit' or obey the rule but cannot make it true. But here it does make sense to describe a single case as possibly ‘making a rule false’.

Just to make things yet muddier, there is arguably a third category of rules which express ‘statistical’ tendencies where an example can contribute a small amount of information to the degree of the tendency of a rule, but no more or less. “If a man is tall then he tends to be heavy” might be such an example. Here we might talk technically of correlation coefficients and ‘outliers’, and we might reject the truth of the rule if the correlation coefficient turned out to have the wrong sign (or to be too small), but again there is no question of a single case establishing the truth or the falsity of such a rule.

So the relations in the card-to-rule direction are particularly complex. Unlike in the semantically simple deontic cases, the semantics of indicative rules is inherently contextually determined. In almost all cases, no clear meaning is attributed to the idea of a single card ‘giving a truth value’ to the rule, and even in the case where this is so, it is only holds for falsifying cases. And different relations hold between whole sets of cards and rules where those cards exhaust the relevant domain of interpretation. It seems not too outlandish a hypothesis that subjects don’t have a uniform interpretation (either within or between themselves) of these relations. Legal laws stand out as the *only* cases where the semantic relations are stark and simple in both directions, and where the difference between the relations of single cases and sets of cases are not confusable. This hypothesis could explain different subjects’ behaviour in ‘abstract’ and ‘thematic’ tasks as an interaction between rule and task interpretation.

A close relative of this hyposthesis did enjoy an outing in the literature, starting out from Yachenin & Tweeney 1982, specifically focussed on explaining differences between abstract and thematic rules and instructions. They noted that abstract rules were invariably accompanied by the instruction to find out *if the rule was true or false*: deontic rules by the instruction to find whether the rule had been *violated*. The early discussion saw this as an instructional difference, and the associated experimental investigations generally explored the idea through instructional manipulations. Perhaps deontic rules worked because the experiments used instructions to seek violations, and this focusses attention on ‘falsification’. Perhaps violation instructions would produce falsificatory behaviour with abstract rules? The ensuing experiments established that instructional manipulations alone (e.g. telling subjects presented with an abstract rule to turn cards that might violate it) did not lead to large increases in turning the negation of the consequent card. Only when the instruction to seek violation was combined with a deontic rule; with a ‘reduced array selection task’ which presented only consequent-visible cards; or with the extra task of providing verbal justifications, did it increase the turning of $\neg q$.

What is proposed here is that the instructional differences between descriptive and deontic tasks are correlated with contrasts in interpretation of the semantic relations between cards and rule, and it is therefore not so suprising that merely changing the instruction wording without giving clearer signals about the intended interpretation of the rule, does not change performance greatly.

What support for this hypothesis can we marshall, and what further evidence could be sought? There is considerable evidence already available that confusion reigns among subjects about the alternative semantic relations between rule and cards, cards and rule, and whether the relation is between sets of cards or singleton cards. The most immediate evidence comes from thinking aloud protocols and the justifications subjects give. Subjects frequently exhibit confusion about whether they are to give judgements about whether the rule is true (or false) of a single card, or of the whole set, often revealed by such statements as 'well whether I turn this depends on what's under that one'. When subjects view the semantic relation from the opposite end, asking themselves about what a card says about a rule, they show even more confusion about what a single card can 'say about a rule'—and justifiably so. It simply isn't clear that subjects assume that the rule is false if it is false of a single card, and this is a reasonable assumption to question if a there is a range of law-likeness of interpretations for descriptive rules.

Another kind of evidence comes from construction and evaluation tasks. The descriptive abstract rules used reveal what is usually referred to as 'the defective truth table'—a table that values both $\neg p$ plus q, and $\neg p$ plus $\neg q$ cards as 'irrelevant' to the rule *if p then q*. This concept of irrelevance figures prominently in subjects' verbal justifications and think-aloud-protocols with descriptive rules. There are good logical grounds for rejecting this as the *truth* table of 'if ... then', but perhaps it is some other kind of table? A 'compliance' table perhaps?

We do not know of any experiment eliciting what might be called the corresponding 'compliance tables' for deontic conditionals, but we predict the following. For a deontic law such as "if p then must q", P plus Q, $\neg p$ plus q, and $\neg p$ plus $\neg q$, all comply, whereas p plus $\neg q$ violates. The over age drinker, the over age teetotaller, and the under age teetoller all comply with the law (which does not mean their behaviour is controlled by the law). Only under age drinkers are violaters. Laws are *relevant* to all within their domain of aplication. Compliance tables are not defective. The issue of whether observation of a compliant over age teetotaller 'makes the law apply' never arises. No cases ever make the law apply. The direction of the semantic relation is quite unambiguous. Confusion about whether the task requires judgements of what single cards vs sets of cards 'say about the rule' is unlikely ever to arise in deontic cases, since it is absolutely clear that whether one citizen is in violation the drinking age law is quite independent of whether another is (conspiracies to drink being violations of a different law!).

Two further pieces of evidence that this is a promising area in which to look for explanations come from deontic rules embedded in 'descriptive tasks', and experiments where law-like interpretations of decriptive rules are contextually encouraged. We have seen that when Gigerenzer & Hug's instructions to descriminate which of two laws exists (the hikers or the guides must bring the wood), this descriptive task depresses performance, presumably because it complicates semantic relations between rule, cards and isntructions.

Secondly, when Almor & Sloman use descriptive rules which are most readily interpreted as based on qualitative natural laws, a descriptive task gets high levels of normative performance. This last result suggests that a natural law-like interpretation yields some of the semantic simplicity of a legal law. Subjects can easily decide which cards appear to violate the law, no matter if violation of background conditions is in fact due to mundane human interference.

In summary, the semantic relations between rule and card, card and rule, rule and sets of cards, and sets of cards and rule are all distinct but highly confusable in the descriptive case, at least without further context. They are completely unambiguous in the deontic case. Merely changing to violation instructions for a descriptive rule without giving other clues about the required interpretation is unlikely to solve the problems.

Far more needs to be done to show that the contextual determination of the form of interpretation of task and rule systematically determines subjects' behaviour and justification. At this stage we believe that a return to the strategy of analysing subjects' dialogues with the experimenter is a promising avenue to yield information on the instability of subjects' interpretations during reasoning. We already have a corpus of such dialogues rich in anecdotal illustration of subjects' struggles to make explicit their semantic theories. To go beyond anecdote requires a systematic framework for aligning utterance with behaviour. However fraught that course is known to be, it has two great recommendations. It promises to found psychological theory on semantics, thus bridging laboratory speak with the vernacular; and it is highly suggestive of an approach to teaching some important fundamentals of reasoning and communication.

8 Conclusions

Our analysis notes that meaning acts as a hidden variable in psychological theories of reasoning, a variable which is not controlled for but which may influence the outcome in unforeseen ways. Although semantics is far from a finished science, and the details of accounts of conditionals and tasks will vary, the issues that must be resolved have been well known in the semantics literature for some time. The semantic relations between rules and cases are various and complex. Truth, relevance, compliance are different relations. They are asymmetrical relations. This much is uncontroversial and it is all we need to argue for the necessity firmer semantic foundations and more articulated explanations of interpretation, reasoning, behaviour and experience.

But to turn our proposals into an empirical theory we need evidence about the actual mental processes involved (at several levels); how these differ with different rules, tasks, contexts and reasoners; and how they change in the process of learning.

We believe that answering these questions calls for much richer empirical evidence than has typically been the focus of experiments.

Ideally, one would like to control meaning completely; if that were possible, one could then proceed in the conventional experimental course. However, as we have argued, complexity of reasoning influences interpretation, and so full control appears to be impossible.

The next option would be to design the experiment in the form of structured interviews, where one tries to elicit the meaning an subject assigns to an important expression as he performs a particular task. Such a design is definitely more vulnerable to subjective interpretations than the standard designs, but if carried out and analysed carefully, may yield more insight into the interplay between meaning and reasoning. Finding out what has to be done to teach insight should be viewed as another kind of evidence about the initial state of the student. Though verbalisations may in some circumstances be quite tangentially related to the behaviour they 'rationalise', education is a process which can bring these processes into closer alignment. We believe that the teaching of logic would benefit greatly from a careful study along these lines, because it would give a clearer idea of the student's baseline, from which all instruction has to start. Such a study would also contribute to our understanding of the relation between behaviour and phenomenology.

We hope at least, that even if our readers won't agree whether psychology is hard or impossible, that they will at least agree with the presupposition that it is not easy?

References

[1] A. Almor and S.A. Sloman. Is deontic reasoning special? *Psychological Review*, 103(2):374–380, 1996.

[2] J.R. Anderson. *The adaptive character of thought.* Lawrence Erlbaum Associates, Hillsdale, NJ., 1990.

[3] R. J. Bracewell and S. E. Hidi. The solution of an inferential problem as a function of stimulus materials. *Quarterly Journal of Experimental Psychology*, 26:480–488, 1974.

[4] K. Cheng, P.and Holyoak. Pragmatic reasoning schemas. *Cognitive Psychology*, 14, 1985.

[5] L. Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? studies with the Wason selection task. *Cognition*, 31:187–276, 1989.

[6] E. E. Vallduví, E. Engdahl. The linguistic realization of information packaging. *Linguistics*, 34(3):459–519, 1996.

[7] J.St.B.T. Evans, S.L. Newstead, and R.M. Byrne. *Human reasoning: the pychology of deduction.* Lawrence Erlbaum Associates, Hove, Sussex, 1993.

[8] J.St.B.T. Evans and D.E. Over. Rationality in the selection task: epistemic utility versus uncertainty reduction. *Psychological Review*, 103(2):356–363, 1996.

[9] S.I. Fillenbaum. How to do some things with if. In Cotton and Klatzky, editors, *Semantic functions in cognition.* Lawrence Erlbaum Associates, 1978.

[10] G. Gebauer and D. Laming. Rational choices in wason's selection task. *Psychological Research*, 60:284–293, 1997.

[11] M. C. Geis and A. M. Zwicky. On invited inferences. *Linguistic Enquiry*, 2:561–566, 1971.

[12] G. Gigerenzer and K. Hug. Domain-specific reasoning: social contracts, cheating, and perspective change. *Cognition*, 43:127–171, 1992.

[13] R. A. Griggs. Memory cueing in instructional effects on wason's selection task. *Current Psychological Research and Review*, 3:3–10, 1984.

[14] R. A. Griggs and J. R. Cox. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 73:407–420, 1982.

[15] M. Henle. On the relation between logic and thinking. *Psychological Review*, 69:366–378, 1962.

[16] P. N. Johnson-Laird and R.M. Byrne. *Deduction.* Lawrence Erlbaum Associates, Hove, Sussex., 1991.

[17] A.I. Khinchin. The entropy concept in probability theory. In *Mathematical foundations of information theory.*, pages 2–28. Dover, 1957.

[18] D. Laming. On the analysis of irrational data selection: a critique of oaksford and chater. *Psychological Review*, 103(2):364–373, 1996.

[19] K. I. Manktelow and J.St.B.T. Evans. Facilitation of reasoning by realism: effect or non-effect? *British Journal of Psychology*, 71:227–231, 1979.

[20] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman, San Fransisco, 1982.

[21] M. R. Oaksford and K. Stenning. Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, memory and cognition*, 18:835–854, 1992.

[22] M.R. Oaksford and N.C. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101:608–631, 1994.

[23] M.R. Oaksford and N.C. Chater. Rational explanation of the selection task. *Psychological Review*, 103(2):381–392, 1996.

[24] D. Sperber, F. Cara, and V. Girotto. Relevance theory explains the selection task. *Cognition*, 57:31–95, 1995.

[25] D. Sperber and D.. Wilson. *Relevance: Communication and Cognition.* Blackwell, Oxford, 1986.

[26] K. Stenning, R. Cox, and J. Oberlander. Attitudes to logical independence: traits in quantifier interpretation. In *Proceedings of Seventeenth Meeting of the Cognitive Science Society, Pittsburgh 1995.*, pages 742–747. 1995.

[27] P. C. Wason. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behaviour*, 4:7–11, 1965.

[28] P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20:273–281, 1968.

[29] P. C. Wason and P. N. Johnson-Laird. A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology*, 61(4):509–515, 1970.

[30] P. C. Wason and P. N. Johnson-Laird. *Psychology of Reasoning: Structure and Content.* Harvard University Press, Boston, 1972.

[31] P. C. Wason and D. Shapiro. Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23:63–71, 1971.