Semantics as a foundation for psychology: a case study of Wason's selection task.*

Keith Stenning	Michiel van Lambalgen
HCRC	ILLC
University of Edinburgh	University of Amsterdam
k.stenning@ed.ac.uk	michiell@wins.uva.nl

February 17, 2000

Abstract

We review the various explanations that have been offered to account for subjects' behaviour in Wason's famous selection task. We argue that one element that is lacking is a good understanding of subjects' 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. We make novel proposals in these terms for explaining the major content effects of deontic materials. Throughout we illustrate with excerpts from tutorial dialogues which motivate the kinds of analysis proposed. Our long term goal is an integration of the various insights about conditional reasoning on offer from different cognitive science methodologies. The purpose of this paper is to try to draw the attention of logicians and semanticists to this area, since we believe that empirical investigation of the cognitive processes involved could benefit from semantic analyses.

1 Introduction

When Peter Wason invented his '4-card' task (e.g. Wason [40]), he created one of cognitive science's fruit flies—a laboratory phenomenon of decep-

*The research for this paper was partially supported by ESRC Fellowship (GR R34938) to the first author; by EPSRC visiting fellowship (GR 34657) to the second author; and by ESRC Centre Grant (GR M423284002) to HCRC. The second author is also indebted to the Netherlands Organisation for Scientific Research (NWO) for support (grant PGS 22 262). We are deeply grateful to Magda Osman for doing part of the experiment, transcribing the interviews, for helpful comment and doing the statistical analyses.

tive simplicity which is a potential basis for theory reaching 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. One conclusion of our analyses is that an integrated account of the many phenomena involved urgently requires a broadening of the empirical evidence base. We illustrate our analyses throughout with excerpts from tutorial dialogues. The semantic analysis proposed here provides the basis for guiding an enrichened experimental program.

2 Task and Phenomena

Wason's task in its original form [40] involves the choice of evidence relevant to the truth or falsity 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 (this part will henceforth be referred to as the' background rule'). A conditional rule (henceforth: 'foreground rule') is then presented which might or might not be true of 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 instructions state that the rule applies only to the four cards shown. The reasoner's task is to turn those cards and only those cards which it is necessary to turn in order to determine whether the rule is true or false. Four cards bearing on their visible face, say, 'A', 'K', '4' and '7' appear below the rule.

In this and many subsequent replications, populations of intelligent undergraduate students 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. If the rule is formalised as $p \rightarrow q$, a typical distribution of results is as follows: p, q 46 %, p 33%, $p, q, \neg q$ 7%, $p, \neg q$ 4% and others 10% (from Wason and Johnson-Laird [42]).

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. Performance in both construction and evaluation tasks generally accords well with the classical logical competence with the exception that cases with false antecedents are often evaluated as 'irrelevant'. Specifically, almost all subjects *evaluate* a 7/A card as falsifying the conditional. The selection task clearly involves something more than simply conditional reasoning.

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 work ([14] and later) is a worthy exception, but perhaps one that proves the rule. There has been some linguistic interest (Geis and Zwicky [16]) in the relation between the psychological observations and the theory of pragmatics. Philosophical work on the Ravens Paradox (see below, section 9) 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 [43]; Johnson-Laird and Byrne [23]). This response has especially been engendered by what are known as *thematic* or *content* effects. Early after Wason's initial experiment, Wason and Shapiro [44]) and Wason and Johnson-Laird [43] 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 early 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, as even the casual reader of volumes such as *On conditionals* [37] and *On conditionals again* [5] will have noticed.

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 [27]). 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 [19] showed that regulations provided particularly facilitating kinds of thematic linkage. Cheng and Holyoak [7] 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 [10]), and this for evolutionary reasons. In this context 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, and deontics have many uses other than expressing social obligations. The 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 [17]). 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)). At this point the reader may have noticed that the instruction in these deontic tasks, which refer to 'obeying the rule' are subtly different from those in the original task, which refer to the truth value of the rule. This difference will be of some importance below.

There have been claims to the effect that good performance can also be achieved without resorting to deontic material and particular social perspectives. Wason and Green [41] found identical performance for a plausible drinking age rule, an absurd drinking age rule, and an indicative conditional stating an arbitrary relation between colour and lengths of bits of wool described as a quality control regularity in a factory. They used a reduced array selection task in which subjects were only offered consequent cards (i.e. $q, \neg q$). With this task, the probability of turning the false consequent card is much higher, but the point here is that the three conditionals elicitd identical performance. If Wason and Green's third rule is claimed to be a social contract, the concept of social contract has been so extended as to become meaningless.

Sperber et al. [32] 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 [18] earlier failed to find such facilitation. This can be explained by the fact that they used an 'epistemic' perspective on their rule, of the kind Gigernzer and Hug showed to be relatively ineffective. Sperber et al.'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.

3 What has to be explained?

If these are the bare outlines of some salient 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? For that matter, do the data outlined above say all there is to say about performance? How does a cognitive theory of performance in this task relate to other areas of cognitive science? 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. Minimally, understanding conditional rules requires recognition of word forms and syntactic structures. Distinguishing social obligations from descriptions is a formal classification, presumably triggered by complex contextual features.

Here are two examples of what we mean by form, the first negative, the second positive.

Psychologists (cf. Griggs and Cox [19]) have sometimes been tempted to speculate that people can only reason with conditionals which they can remember from past experience. This so called memory cueing would lead to good performance in those cases where subjects are familiar with counterexamples, and only in those cases. At one point, observing that Plymouth undergraduates performed differently with the transport problem than their London peers, Newstead speculates (in *op. cit.*) that this could be explained by supposing that they can retrieve reasoning about one transport destination from memory but not another. Even such a 'memory' theory of conditional reasoning must assign some role to form (in virtue of which retrieval takes place), but will, needless to say have redoubtable problems with understanding the processes of cognitive development and of transfer of reasoning precisely because of its minimal appeal to form.

For a positive example, consider Comrie's paper [9], where he argues that different types of conditionals are distinguished by the degree of hypotheticality of their antecedents. Although he believes that no conditional entails the truth of its antecedent, the degree of hypotheticality may vary on a continuum from very likely, grammatically marked by indicative mood without backshifting in tense, to highly unlikely or even counterfactual, grammatically marked by the pluperfect. Athanasiadou and Dirven [4] partly disagree and point to the 'course-of-event' conditional which states the regular co-occurence of two events which are in a relation of dependency; here the truth of antecedent and consequent seems to be determine a value of this parameter, but the response patterns of subjects in Wason's task may well be affected by an attempt to do so. One could believe that many such parameters which are determined in everyday communication are simply irrelevant to the laboratory behaviour. But if so, so much the worse for the laboratory.

More generally, 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 assignment plays in determining reasoning and choice. We would like to connect theories of the forms of sentences to theories of reasoning with sentences. The general picture is this: in order to solve the abstract task, the subject may need to set a number of parameters, parameters which are already set by the context in the thematic tasks. To investigate what the possible parameter settings are, we have to collect data of different type, for instance on the pragmatic inferences from conditionals considered by Fillenbaum [14]. Somewhat surprisingly, subjects also differ in their interpretations of semantic notions such as 'true', 'false', 'obey', 'satisfy' or 'fit', and this is also clearly an aspect of the form assigned to the task.

This naturally leads to one feature of this requirement 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 shows a strong tendency for the same behaviour to be repeated (see, for example, Gebauer and Laming [15]). 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. As a concrete example, we may mention the p, q choice: we present evidence that this choice is made for such vastly different reasons that it becomes problematical to look for one single explanation.

Wason's early investigations of insight also point to a third desideratum for theory. We would like to understand the subjects' own access to the processes involved in solving the task—what might be termed the *phenomenology* of the task. Running subjects in this task generates 'aha!' experiences (as well as 'oh damn!' experiences). For example, 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. Even those subjects who never attain insight in the sense of giving the normative response, may give lucid explanations for their choice which deserve to be taken seriously. 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. Again, this leads to the consideration of additional data, in the form of tutorial protocols.

Our choices of theoretical aim and empirical method are conditioned by views about the relation between laboratory behaviour, everyday communication, and formal education. As we will see below, we believe laboratory behaviour is frequently superficial because subjects' assimilate the situation to more familiar ones in a variety of unintended ways. However, we do not conclude that learning the intended assimilation is either trivial or educationally unimportant. Understanding the learning processes involved in coming to the highly objectified stance toward language which the task demands, may be a more useful goal for cognitive theory, than treating the superficial stances initially adopted as indicators of 'natural' cognitive mechanisms.

4 Tutorial interviews

A standard 4-card task experiment consists in giving subjects a form which contains the instructions and shows four cards; the subjects then have to mark the cards they want to select. The type of data obtainable in this way is highly abstracted from the reasoning process. The subjects' approach to the task may be superficial in the sense of not engaging any reasoning or comprehension process which would be engaged in plausible real-world communication with the relevant conditionals. One looses information about subjects' vacillations (which can be very marked) and thus one has little idea at what moment of their deliberations subjects make a choice. It is also possible that the same answer may be given for very different reasons. Furthermore, the design implies that the number of acceptable answers is restricted; for instance, some subjects are inclined to give an answer such as 'A or 4', or 'any card', or 'can't say, because it depends on the outcomes', and clearly the standard design yields no information about this. Early on, Wason and Johnson-Laird, in several papers, investigated the relationship between insight and reasoning by also using interviewing protocols. They distinguished two kinds of feedback: 1) feedback from hypothetical turnings - —'suppose there is a A on the back of the 7, what would you then conclude about the rule?'; 2) actual feedback in which the subject turns the 7-card and finds the A—'are you happy that you did/didn't select the 7 card?'. It seems to us that this type of design is much more conducive to obtaining information about the why's and wherefore's of non-normative answers.

Of course, the rich data of tutorial dialogue brings with it its own problems. We do not interpret these dialogues as *reports* of reasoning that went on before the dialogue, let alone as transparent and complete reflections of such preceding thought processes. These dialogues *are* the subjects' reasoning with a tutor during a dialogue. Engaging subjects in dialogue undoubtedly changes their thoughts, and may even invoke learning. The relation between the reasoning processes evoked by the standard way of conducting the task, and the processes reflected in subsequent dialogues is a relation that remains to be clarified.

All forms of data present problems. The hyper 'objective' data of card selection present problems of interpretation. The subjects' degree of engagement in the task is questionable, as we will presently see. This objective data, because it is so impoverished, leads to a focus on trends in group data, but ignores differences between subjects' performance as noise. Richer data on each subject strongly suggests that there are many different thought processes may lead to even the same responses, let alone different responses.

We seek converging data of varied kinds. In order to implement the program outlined in section 3, we conducted a number of tutorial experiments, in which subjects were invited to explain their choices and reasoning processes. Where possible we collected baseline performance in conventional tasks before engaging subjects in dialogue, and compare the relations between data from the two sources. The dialogues presented here suggest a range of more conventional experiments. This kind of data could be collected and coded in sufficient quantities to sustain quantitative analysis, but that is not our purpose here. Our purpose is exploration, and the gathering of sufficient evidence to justify more sustained comparison of accounts at a later date. The kind of evidence presented is particularly weak as negative evidence. The fact that we do not observe something is exceedingly weak evidence that it is not to be found. Think rather of these dialogues as 'existence proofs' that a phenomenon does occur.

The sessions were videotaped and then transcribed. The experiment was performed in two runs, March 1999 (19 subjects) and July 1999 (10 subjects). Since the experiment had the character of a pilot study, we changed some instructions during running, when we felt that they could be made more effective. These changes are distinguished in the descriptions that follow. We now give an overview of the conditions.

1. Inducing correct understanding of anaphoric expression 'one side - – other side' by treating the three possibilities explicitly

    (a) if there is a vowel on the (visible) face, then there is an even number on the (invisible) back

    (b) if there is a vowel on the (invisible) back, then there is an even number on the (visible) face

    (c) if there is a vowel on one side (face or back), then there is an even number on the other side (face or back)

    In all cases the cards shown were AK47. The A carried a 7 on the back, the K a 4, the 4 a K and lastly the 7 an A. In the first two conditions we asked subjects to select the cards. The last condition was introduced by explaining that the first two conditions did *not* represent the intended meaning of the anaphora, but that the intended reading is symmetric with respect to the sides of the card. This tutoring was intended to have the effect that the cards were taken to be reversible—we will see evidence of how effective the intervention was. At each phase of tutoring, we first asked a subject to imagine what could be on the invisible side of a card, what that would mean for the rule, and we then proceeded to the actual turning of all the cards. At the end we asked subjects whether they were happy with their original selection. This condition was included only in the first run. These conditions will be referred to as experiments 1a, 1b, 1c respectively.

2. Investigating the role of 'task semantics' by providing an instruction reminiscent of the one used for deontic tasks. The rule is still 'if there is a vowel on one side, then there is an even number on the other side', but the subject is now asked to select those cards which have to be turned in order to determine whether they *obey* the rule. The cards shown were EG25. E carried 5 on the back, G 2, 2 G and lastly the 5 carried an E. We first asked subjects to select cards, then to imagine what could be on the other side, and lastly to turn all cards. At the end we gave subjects the opportunity to revise their earlier selection. This condition will be referred to as experiment 2.

3. A novel two-rule task, where subjects were instructed that one is true and the other false, and are asked to decide which is which

   (a) if there is a U on one side, then there is an 8 on the other side
   (b) if there is an I on one side, then there is an 8 on the other side

   given that one side contains U or I, the other side contains 3 or 8. The cards shown were UI83. In this case, both U and I carried an 8, 8 carried an I, and 3 a U. Again, we first asked subjects to select cards, then to imagine what could be on the other side, and lastly to turn all cards, after which subjects were given the opportunity to revise their earlier selection. Although we ran this condition on 29 subjects, we will use only data from subjects 20–29, since the instruction sheet for the other subjects contained a mistake. The purpose of this condition was to encourage the subjects to adopt a stance for which a single counterexample would be sufficient to falsify a rule (cf. section 10). Furthermore, bearing in mind the Bayesian explanation of performance in terms of a hidden alternative rule (cf. section 9), we were interested how subjects performed with an alternative explicitly given. This condition will be referred to as experiment 3.

4. During the same session we gave subjects a booklet to fill in, which consisted of two parts: one part containing the selection tasks outlined above, the other part containing a number of sentences logically equivalent to the conditional at issue - here subjects were asked to select

paraphrases from a given set, or to supply their own. For example, the sentence given could be 'it is not the case that there is a vowel on one side and an odd number on the other side', and among the paraphrases provided there were sentences like 'if there is a vowel on one side, then there is an even number on the other side'. Some subjects received the booklet before the interview, in order to get a baseline for their performance.

The remainder of the paper discusses existing explanations in the light of our observations. Specifically we will illustrate theories proposed by means of dialogue fragments, give cases which clearly do not fit a proposed theory, and highlight phenomena which do not fit into any of the existing theories.

Just to give the reader a preliminary idea of what such dialogues can look like, we give two excerpts which illustrate phenomena also noticed by Wason and Johnson-Laird.

The first example shows that (in experiment 1c) subjects may fail to understand the implication of the 7/A combination. Here, as in the sequel, we denote by '7/A' the card which has 7 on the visible face and A on the invisible back.

Example Subject 14 [experiment 1c]

> *S.* I would just be interested in A's and 4's, couldn't be more than that.
> *E.* So now let's turn the cards, starting from right to left. [Subject turns 7 to find A] Your comments? *S.* It could be an A, but it could be something else ...
> *E.* So what does this tell you about the rule?
> *S.* About the rule ... that if there is an A then maybe there is a 7 on the other side.
> *E.* So there was a 7.
> *S.* But it doesn't affect the rule.

The second example shows that a subject sometimes hypothesises (or discovers) an E on the back of 5, 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).

Example Subject 3 [experiment 2]

> *E.* OK Lastly the 5.
> *S.* Well I wouldn't pick it.
> *E.* But what would it mean if you did?
> *S.* Well, if there is an E then that would make the rule false, and if there was a G, it wouldn't make any difference to the rule.

Wason and Johnson-Laird report that subjects can normatively justify card choices when those choices are presented to them (rather than elicited from them). In fact, as the evaluation and construction tasks have shown, *reasoning* about the cards does not always seem to be the problem; and we see in the above excerpt that the subject adequately judges the import of the 5/E and 5/G cards. Rather it seems to be the interplay between reasoning and selection that causes trouble. We shall have much more to say on this issue below.

We now turn to the various explanations that have been provided of performance in the 4-card task. We will discuss these under the following headings: *matching bias*, *non-standard interpretation*, *verification bias*, *social contract theory*, *Bayesian* and *task semantic* explanations. 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.

5 Matching Bias

Evans (see for example the review in Evans, Newstead and Byrne [12]) 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*. Here is a particularly striking

Example Subject 9 [experiment 1a]

> *E.* [This rule] says that if there is a vowel on the face, then there is an even number on the back. So what we mean by face is the bit you can see, and by back the bit you can't see. Which cards would you need to turn over to check if the rule holds?
> *S.* This one [ticks A] and this one [ticks 7]
> *E.* So why would you pick those two?
> *S.* One has vowel on the face and the other one an even number. If you turn it, if it's true then it should have an even number [pointing to the A] and this should have a vowel[pointing to the 7].
> *E.* [baffled] So you picked, oh you were saying if there was a vowel underneath [pointing to the 7]
> *S.* That's because I'm stupid. Even number is 1,3,5, ...
> *E.* No, 2,4,6, ...
> *S.* [Corrects 7 to 4, so her final choice was A and 4] OK So these.

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 [28] by investigating a full range of clause negations in both selection and evaluation tasks, showed that matching is not a particularly 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. In fact, this was one of the purposes of experiment 3, where, interestingly, no subject used 'matching' terminology to justify their choice. *A priori*, One could think that there are two kinds of matching responses: {U,I,8} and {8}. The first one didn't occur, and the second choice was invariably motivated by arguing that either rule one or rule two was true, depending on whether there is U or I on the back of 8. For more on this type of argument, see section 6.5. Subject 21 chose 8,3, explicitly noting that 3 was not mentioned in the rule.

6 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 [21] 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 [40] considered the possibility of a 'biconditional' interpretation of the conditional, and Bracewell and Hidi [6] proposed that the 'one side - other side' anaphor in the rule might be interpreted in a constant rather than a variable reading. We shall see, however, that there are other, more subtle, possibilities for nonstandard interpretations, not only of key terms in the *rule* ('if – then', 'one side–other side'), but also of key terms in the *instructions* ('true', 'false', 'obey'). Furthermore, these interpretations may interact.

6.1 Anaphora

The most plausible 'constant' reading of the anaphor 'one side - other side' results in an interpretation which can be paraphrased: 'if there's vowel on the (visible) face of the card, then there's an even number on the (invisible) back.' Adopting this interpretation (along with a conditional rather than biconditional reading) would explain subjects' choosing just the p card. (For another explanation of this choice, see below, section 6.5.) Similarly, adopting this interpretation together with a biconditional reading could explain the selection of the p, q cards. Johnson-Laird and Wason [24] referred to this phenomenon by saying that subjects do not always recognise the *reversibility* of the cards. In another paper, Wason and Johnson-Laird [42] tried to eliminate this factor by working with cards where all information was present on one side, and where some of the information was masked; subjects were then asked to select those cards which had to be unmasked. The results did not differ significantly from the pattern of answers in the standard task. This could be explained in two ways, not mutually incompatible. Firstly, it is not so much the asymmetry between face and back, as the asymmetry between known and unknown, that is operative here. Secondly, the intended reading of the anaphora remains computationally difficult also in the modified design, because the referent of 'other part' depends on the referent for 'one part'; it is precisely this dependence that is eliminated in the constant reading, where 'one side' refers to 'face' (or known information) and 'other side' refers to 'back' (or unknown information). In other words, only on the intended reading is 'other side' a real anaphor, whose referent is however not given directly by the antecedent of the conditional, but has to be computed.

More recently Gebauer and Laming [15] 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.[1]

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 and 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. In fact, we have a number of examples which show that subjects do not independently adopt interpretations for the anaphora and for the conditional; rather, there can be influence both ways. We shall give some examples after we have treated the biconditional interpretation more fully. To conclude our discussion of Gebauer and Laming, we briefly discuss a classroom experiment performed by us in May 1998, designed to test whether subjects are sensitive to explicitly given anaphoric relationships.

We gave 81 subjects one rule each from four different formulations of the rule:

1. if there is a vowel on one side of the card, then there is an even number on the other side

2. if there is a vowel on one side of the card (face or back), then there is an even number on the other side (face or back)

3. if there is a vowel on the face of the card, then there is an even number on the back

4. if there is a vowel on the back of the card, then there is an even number on the face.

[1] Four combinations, because the constant back/face reading of the anaphor appears to be too implausible to be considered.

The data are presented in Appendix A, section 12. Somewhat surprisingly, there were no significant differences between the conditions, and the answers followed the standard pattern. In fact they were statistically indistinguishable from Wason's original data. This population of subjects (Edinburgh first year introductory psychology course students) has been used in replicating selection task, and many other standard results in the reasoning literature. Only three of eighteen subjects presented with rule 4 responded without turning a letter card. Normative choice for Rule 4 is to choose just the 7. Subjects' choices were indistinguishable from choices for the three other rules. Students are not processing the difference between 'the face' and the 'the back'. The processing that goes on is grossly insensitive to the different wordings of the rule. Although running a very large sample might reveal a few subjects who are reading closely, the power of the experiment was sufficient to expect identifiable effects for such grossly different rules. The condition samples sizes were comparable to those on which this literature is based.

Furthermore, there is no significant difference between this classroom experiment's results, and the corresponding conditions administered one-on-one by the tutors prior to tutoring. Deeper processing is invoked by interactive dialogue, but not by the difference between classroom and one-on-one task administration.

This seems to argue against Gebauer and Laming's suggestion that subjects have definite, although different, interpretations of the anaphora, at least if these interpretations are supposed to be related to interpretations of the English sentences that might arise outside this task. We return to the interactions between reasoning and interpretation in section 6.3. This was one reason why we decided to try tutorial interviews; these might encourage subjects to think more deeply about the meaning of the key terms.

6.2 Conditional as biconditional

Geiss and Zwicky [16] 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: if I would buy you lunch in any case, my conditional promise would be pointless. On the other hand, for non-deontic conditionals, the 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.

Although some deontics encourage biconditionality, the legal law ones used as thematic material in this literature are not prone to these interpretations, probably because closed-world assumptions are most unlikely. From " If you are under 18, then you mustn't drink alchohol" we are most unlikely to conclude that "If you mustn't drink alchohol, then you are under 18" because we know that there are lots of other reasons (such as driving) for abstinence.

On its own, a purely interpretational hypothesis of the form 'conditional = biconditional' would suggest that a subject interpreting the rule biconditionally would turn all four cards. This is a rather rare event. Such a hypothesis by itself hardly helps to explain the modal choice of just the p and q cards; additional hypotheses, for example that the subjects adopts a constant reading for the anaphora, are necessary. Below we discuss some further possibilities.

6.3 New findings: interference effects

As promised, we now provide a number of examples which demonstrate the interplay between the interpretations chosen for anaphora and conditional. The first example shows us a subject who explicitly changes the direction of the implication when considering the back/face anaphora, even though she is at first very well aware that the rule is not biconditional.

Example Subject 12.[experiments 1a,1b,1c]

> *E.* The first rule says that if there is a vowel on the face of the card, so what we mean by face is the bit you can see, then there is an even number on the back of the card, so that's the bit you can't see. So which cards would you turn over to check the rule.
> *S.* Well, I just thought 4, but then it doesn't necessarily say that if there is a 4 that there is a vowel underneath. So the A.
> *E.* For this one it's the reverse, so it says if there is a vowel on the back, so the bit you can't see, there is an even number on the face; so in this sense which ones would you pick?

S. [Subject ticks 4] This one.
E. So why wouldn't you pick any of the other cards?
S. Because it says that if there is an even number on the face, then there is a vowel, so it would have to be one of those [referring to the numbers].

⋮

E. [This rule] says that if there is a vowel on one side of the card, either face or back, then there is an even number on the other side, either face or back.
S. I would pick that one [the A] and that one [the 4].
E. So why?
S. Because it would show me that if I turned that [pointing to the 4] over and there was an A then the 4 is true, so I would turn it over. Oh, I don't know. This is confusing me now because I know it goes only one way.

⋮

S. No, I got it wrong didn't I, it is one way, so it's not necessarily that if there is an even number then there is a vowel.

The second example is of a subject who gives the normative response in experiment 1c, but nonetheless goes astray when forced to consider the back/face interpretation.

Example Subject 4.[experiments 1a,1b,1c]

E. OK This says that if there is a vowel on the face [pointing to the face] of the card, then there is an even number on the back of the card. How is that different to ...
S. Yes, it's different because the sides are unidirectional.
E. So would you pick different cards?
S. If there is a vowel on the face . . . I think I would pick the A.
E. And for this one? [referring to the second statement] This is different again because it says if there is a vowel on the back . . .
S.[completes sentence] then there is an even number on the face. I think I need to turn over the 4 and the 7. Just to see if it (the 4) has an A on the back.
E. OK Why wouldn't you pick the rest of the cards?
S. I'm not sure, I haven't made up my mind yet. This one (the A) I don't have to turn over because it's not a vowel on the back, and the K is going to have a number on the back so that's irrelevant. This one [the 4] has to have a vowel on the back otherwise the rule is untrue. I still haven't made up my mind about this one (the 7). Yes, I do have to turn it over because if it has a vowel on the back then it would make the rule untrue. So I think I will turn it over. I could be wrong.

> [When presented with the rule where the anaphora have the intended interpretation]
> *S.* I would turn over this one (the A) to see if there is an even number on the back and this one (the 7) to see if there was a vowel on the back.

Our third example is of a subject who explicitly states that the meaning of the implication must change when considering back/face anaphora.

Example Subject 16. [experiments 1a,1b,1c]

> [Subject has correctly chosen A in first anaphora condition.]
> *E.* The next one says that if there is a vowel on the back of the card, so that's the bit you can't see, then there is an even number on the face of the card, so that's the bit you can see; so that again is slightly different, the reverse, so what would you do?
> *S.* Again I'd turn the 4 so that would be proof but not ultimate proof but some proof . . .
> *E.* With a similar reasoning as before?
> *S.* Yes, I'm pretty sure what you are after . . . I think it is a bit more complicated this time, with the vowel on the back of the card and the even number, that suggests that if and only if there is an even number there can be a vowel, I think I'd turn others just to see if there was a vowel, so I think I'd turn the 7 as well.
> [In third condition chooses A and 4]

So far the examples have been concerned with the influence of the reading adopted for the anaphora on the interpretation of the conditional. We now present an example which shows that the influence can go both ways.

Example Subject 23 [experiment 2]

> *S.* Then for this card [2/G] the statement is not true.
> *E.* Could you give a reason why it is not?
> Well, I guess this also assumes that the statement is reversible, and if it becomes the reverse, then instead of saying if there is an E on one side, there is a 2 on the other side, it's like saying if there was a 2 on one side, then there is an E on the other.

⋮

> *E.* Now we'll discuss the issue of symmetry, you said you took this to be symmetrical.
> *S.* Well, actually it's effectively symmetrical because you've got this either exposed or hidden clause, for each part of the statement. So it's basically symmetrical.
> *E.* But there are two levels of symmetry involved here. One level is the symmetry between visible face and invisible back, and the other

aspect of symmetry is involved with the direction of the statement 'if ... then'.
*S.*Right, o.k. so I guess in terms of the 'if ... then' it is not symmetrical ... In that case you do not need that one [2], you just need *E.*
[experiment 3; while attempting the task he makes some notes which indicate that he is still aware of the symmetry of the cards] *S.* For U, if there is an 8 on the other side, then rule one is true, and you'd assume that rule two is false. And with I, if you have an 8, then rule one is false and rule two is true.
[The subject has turned the U and I cards, which both carry 8 on the back, and proceeds to turn the 3 and 8 cards.]
S. Now the 3, it's a U and it's irrelevant because there is no reverse of the rules. And the 8, it's an I and again it's irrelevant because there is no reverse of the rules. ... Well, my conclusion is that the framework is wrong. I suppose rules one and two really hold for the cards.
E. We are definitely convinced only one rule is true ...
S. Well ... say you again apply the rules, yes you could apply the rules again in a second stab for these cards [3 and 8] here.
E. What do you mean by 'in a second stab'?
S. Well I was kind of assuming before you could only look at the cards once based on what side was currently shown to you. ... This one here [8] in the previous stab was irrelevant, because it would be equivalent to the reverse side when applied to this rule, I guess now we can actually turn it over and find the 8 leads to I, and you can go to this card again [3], now we turn it over and we apply this rule again and the U does not lead to an 8 here. So if you can repeat turns rule two is true for all the cards.
E. You first thought this card [3] irrelevant.
S. Well it's irrelevant if you can give only one turn of the card.

What's interesting in this exchange is that in the first experiment the variable, 'symmetric' reading of the anaphora seems to trigger a symmetric reading of the implication, whereas in the second experiment asymmetric readings of the anaphora and the implications are conjoined, even though he was at first aware that the intended reading of the anaphora is symmetric. (The fact that the subjects wants to turn the cards twice is evidence for the constant (asymmetric) reading of the anaphora.)

Note that the first experiment tutored the subject to read the implication unidirectionally; as a consequence of this successful tutoring he now also seems to take the anaphora asymmetrically.

The upshot of these examples seems to be that it is too simplistic to impute a fixed interpretation of the rule to the subject, an interpretation which may or may not differ from the one intended by the experimenter.

Rather, the interpretation may be constructed during the execution of the task, and can be very much a dynamic affair. Even subjects who are capable of giving the normative answer show such interference effects. This finding raises a number of questions, for instance: if the interpretation is not fixed, what is one actually testing? what causes the interference effects? can interference effects be used to explain the modal response p, q?

In answer to the first question, the present findings suggest that it is not so much the selections themselves which are of most interest, as the representations constructed in the course of solving the problem.

As regards the second question, there appear to be several possibilities, ranging from working memory effects to semantics, although this is mostly a matter of speculation. A working memory explanation would argue that the anaphora and the implication are represented (either spatially, say as arrows, or verbally, as say sequences) and that it is more difficult to simultaneously remember two different directions (and which direction applies to which concept) than to align them. A semantic explanation could refer to the view that conditionals with consequents known to be true are odd, some would even say ungrammatical (Haiman [20]). In his *Introduction to mathematical logic*, Church gives the following example

> If Hitler was a military genius, London is the capital of England.

Haiman (*op. cit.* argues that such examples violate the prime pragmatic function of conditionals 'if p, q', which is to add the antecedent p to one's stock of beliefs, and then to see whether the consequent q is true[2] Because these conditionals are pragmatically perverse, they may be subject to what Fillenbaum [14] called 'pragmatic normalisation', the process which transforms the threat 'Stop screaming or I won't break your arm' (often unwittingly) into 'Stop screaming or I will break your arm'. Similarly, pragmatic normalisation could lead to a reversal of the implication, to produce a sentence which now makes pragmatic sense.

So, an 'interference' explanation for the choice of the p, q card would run like this. Suppose subjects decompose the intended variable anaphora reading of 'one side – other side' into 'face/back' and 'back/face', and then proceed to reverse the direction of the implication in the latter case. This would lead to the transition from

> If there is a vowel on one side, then an even number on the other side

[2]Clearly this analysis, modelled on Stalnaker, does not take account of diagnostic reasoning, which assumes q as given and inquires whether p could be a cause.

via

> If there is a vowel on the face, then an even number on the back
> and
> If there is a vowel on the back, then an even number on the face

to

> If there is a vowel on the face, then an even number on the back
> and
> If there is an even on the *face*, then a vowel on the *back* [3].

What speaks in favour of this analysis, is that about one third of our subjects consider the K/4 card to be irrelevant, whereas 4/K is taken to falsify (see 7 for more elaborate discussion), a surprising fact which is however entirely consistent with the analysis proposed here. What seems to speak against it, however, is that some subjects who give the normative answer for the intended reading of the rule, reverse the arrow in case of the 'back/face' anaphora. Furthermore, it is sometimes not entirely clear what subjects mean by 'falsify'; if a subject says that 4/K falsifies (s)he may just as well mean that the even on the face is not *caused* by a vowel on the back. However that may be, notice that the analysis proposed differs from assuming that these subjects have a *fixed* biconditional interpretation for the conditional which they then combine with the constant face/back reading of the anaphora, an analysis proposed by Smalley [31] and Gebauer and Laming [15]. Their 'static' analysis may of course apply to some subjects, but the excerpts presented above seem to require a more dynamic analysis.

Thus far the discussion on anaphora has been concerned with the abstract task. This is for good reason, since the use of deontic material removes the the use of the anaphora! A typical deontic rule is formulated as 'if a hiker stays overnight, he must bring firewood', not as 'if one side of the card bears the name of a hiker, the other side indicates whether he has brought firewood'. Now, to judge from the generally good performance in the deontic case, subjects here experience no difficulty with the reversibility of cards. Could it be that the necessary linguistic processing of the anaphora in the abstract case causes difficulties, whereas in the thematic case a different system takes over? Even if this were so, it would not be the whole story since not all thematic material 'works'. And if deontic materials with anaphora should still prove 'easy', then it raises interesting questions about

[3] As we have seen, at least one subject takes the conditional in the back/face anaphora case to be biconditional. This could be used to justify the $p, q, \neg q$ selection.

what representation 'takes over'—perhaps the contentful representation of what is antecedent condition and what is consequent result? At any event, this issue needs empirical investigation.

With hindsight, one can see that the issue of anaphora was implicitly raised by Wason and Green [41], although their focus is on the distinction between a *unified* and a *disjoint* representation of the stimulus. A unified stimulus is one in which the terms referred to in the conditional cohere in some way (say as properties of the same object, or as figure and ground), whereas in a disjoint stimulus the terms may be properties of different objects, spatially separated.

Wason and Green conjectured that it is disjoint representation which accounts for the difficulty in the selection task. To test the conjecture they conducted three experiments, varying the type of unified representation. Although they use a reduced array selection task (RAST), in which one chooses only between q and $\neg q$, relative performance across their conditions can still be compared.[4]

Their contrasting sentence rule pairs are of great interest, partly because they happen to contain comparisons of rules with and without variable anaphora. There are three relevant experiments numbered 2–4. Exp 2 contrasts unified and disjoint representations without variable anaphora in either, and finds that unified rules are easier. Exp 3 contrasts unified and disjoint representations with the disjoint rule having variable anaphora. Exp 4 contrasts unified and disjoint representations but removes the variable anaphora from the disjoint rule while adding another source of linguistic complexity (an extra tensed verb plus pronominal anaphora) to the unified one.

In the first case (their experiment 2) cards show shapes (triangles, circles) and colours (black, white), and the two sentences considered are
(2a) Whenever they are triangles, they are on black cards.
(2b) Whenever there are triangles below the line, there is black above the line.
That is, in (2a) the stimulus is taken to be unified because it is an instance of figure/ground, whereas in (2b) the stimulus consists of two parts and hence is disjoint. Performance for sentence (2b) was worse than for sentence (2a) (for details see Wason and Green [41], p. 604–607).

[4] A RAST sometimes improves performance, e.g. in case of sentence (1) below. However, our condition 1b almost reduces the task to a RAST because the antecedent can only refer to the back of the cards, although the four kinds of card faces are still visible in our condition. It is of interest to observe that this does not lead to increase in single $\neg q$ choices: 60% q; 35% $p, \neg q$; and 5% $\neg q$ choices.

We would describe the situation slightly differently, in terms of anaphora. Indeed, the experimental set-up is such that for sentence (2b), the lower half of the cards is hidden by a bar, making it analogous to condition 1b with its concrete back/face anaphora, where the object mentioned in the antecedent is hidden. We have seen in section 6.3 that some subjects have difficulties with the intended direction of the conditional in experiment 1b. Sentence (2b) would be the 'difficult half' of the variable-anaphora sentence "Whenever there are triangles on one side of the line, there is black on the other side of the line". Sentence (2a) does not contain location-denoting anaphora. With Wason and Green we would therefore predict that subjects find (2b) more difficult.

In Experiment 3, the sentences contrasted were
(3a) All triangles are red.
(3b) All the cards which have a triangle on one half are red on the other half.
The stimulus for (3a) is unified because it concerns the colour of a shape, whereas it is obviously disjoint for (3b). Again, performance was worse for sentence (3b). In terms of anaphora, sentence (3a) has none, whereas (3b) clearly has variable anaphora, as in condition 1c; hence, if variable anaphora is a source of difficulty, then we should predict worse performance for (3b), as observed.

In motivating Experiment 4, the authors attribute to Johnson-Laird the observation that sentence (3b) is "both longer and linguistically more complex" than (3a), which might account for the difference in performance. We have given a specific content to the linguistic complexity, namely the processing of the quantifiers and variables implicit in the anaphora. Anyhow, in order to compensate for this factor, Wason and Green introduce pronominal (constant) anaphora in their formulation of (3a), which now becomes
(4a) If the figure on the card is a triangle then it has been coloured red, whereas (3b) becomes
(4b) All the triangles have a red patch above them.
The same card stimuli and procedure were used as in the previous case. Neither sentence contains variable anaphora, and so no prediction can be made on that basis. The unified/disjoint distinction remains, with some linguistic complexity differences other than variable anaphora. Again, performance was worse for sentence (4b).

Interestingly, however, performance for sentence (4b), which has no variable anaphora, is much better than for sentence (3b), which has. This suggests that disjoint representation and 'one side–other side' type of anaphora both contribute to complexity, even though, of course, variable anaphora presupposes disjoint representation. Wason and Green write that their results 'are consistent with the notion that in everyday reasoning logical form is intrinsically related to the content in which it is expressed' [41, p. 609]. However, to those of us seeking a theory in terms of the processes of finding form in content, it is obvious that the logical forms involved are all different, once one does not abstract from quantifier structure; e.g. (3b) is more complex than (4b). Hence there would seem to be no reason to abandon the search for an explanation in terms of form. Exactly how form is related to performance remains open; we have suggested a possible mechanism in the case when variable anaphora is present. An account of how unified vs disjoint representation would affect performance must await an account of the representation of attribute binding in working memory (see for example Stenning and Levy [36] for an approach to this question); Wason and Green confess they have little to offer here.

6.4 Implication as conjunction

We now return to the possible interpretations of conditionals and their relevance for subjects' understanding of the task. In the literature on Wason's task only two types are distinguished: the uni-directional material implication, and the biconditional. When one turns to the linguistics literature, the picture is dramatically different. Above, we already alluded to Comrie's paper *Conditionals: a typology* [9], where conditionals are distinguished according to the degree of hypotheticality of the antecedent. Viewed crosslinguistically, this degree ranges from certain, a case where English uses *when* ('when he comes, we'll go out for dinner')[5], via highly unlikely ('if we were to finish this paper on time, we could submit it to the proceedings') to false, the counterfactual. We claim that, in order to understand performance in Wason's task, it is imperative to look into the possible understandings of the conditional that a subject might have, and for this language typology appears to be indispensable. An interesting outcome of typological research is that the conditional ostensibly investigated in Wason's task, the hypothetical conditional, where one does not want to assert the truth of the antecedent, may not even be the most prevalent type of conditional. We include a brief discussion of the paper *Typology of* if-*clauses* by Athanasiadou and Dirven [3] (cf. also [5]) to corroborate this point; afterwards we will connect their analysis to our observations.

In a study of 300 instances of conditionals in the COBUILD corpus [8],

[5]Dutch, however, can also use the conditionals marker 'als' here.

the authors observed that there occurred two main types of conditionals, *course of event* conditionals, and *hypothetical* conditionals. The hypothetical conditionals are roughly the ones familiar from logic; an example is

> If there is no water in your radiator, your engine will overheat immediately. ([8],17)

A characteristic feature of hypothetical conditionals is the events referred to in antecedent and consequent are seen as hypothetical, and the speaker can make use of a whole scale of marked and unmarked attitudes to distance herself from claims concerning likelihood of occurrence. The presence of 'your' is what makes the interpretation more likely to be hypothetical: the antecedent need not ever be true for 'your' car. Furthermore, in paradigmatic cases (temporal and causal conditionals) antecedent and consequent are seen as consecutive. By contrast in course of event conditionals such as

> If students come on Fridays, they get oral practice in Quechua (from [9])

or

> If there is a drought at this time, as so often happens in central Australia, the fertilised egg in the uterus still remains dormant ([8],43)

the events referred to in antecedent and consequent are considered to be generally or occasionally recurring, and they may be simultaneous. Generic expressions such as 'on Fridays' or 'as so often happens ...' tend to force this reading of the conditional. E.g. the first example invokes a scenario in which some students do come on Fridays and some don't, but the ones who do, get oral practice in Quechua. The generic expression 'on Fridays', together with implicit assumptions about student timetables and syllabuses, causes the sentence to have the habitual 'whenever' reading. It is also entailed that some students do come on Fridays, generally. These examples also indicate that course of event conditionals refer to events situated in real time, unlike hypothetical conditionals. It should now be apparent that the logical properties of course of event conditionals are very different from their hypothetical relatives. For example, what is immediately relevant to our concerns is that course of event conditionals refer to a population of cases, whereas hypothetical conditionals may refer to a single case; this *is* relevant, because it has frequently been claimed that subjects interpret the task so that the rule refers to a population of which the four cards shown are only a sample (cf. section 9 below). Interestingly, Athanasiadou and Dirven estimated that about 44% of conditionals in COBUILD are of the course of events variety, as opposed to 37% of the hypothetical variety. Needless to say, these figures should be interpreted with caution, but they lend some plausibility to the claim that subjects may come to the task with a nonintended, yet perfectly viable, understanding of the conditional. We will now discuss the repercussions of this understanding for subjects' card selections.

One of the questions in the logical equivalence task asked subjects to determine which of four statements follow from the rule 'Every card which has a vowel on one side has an even number on the other side'. More than half of our subjects chose the possibility 'It is the case that there is a vowel on one side and an even number on the other side'. Fillenbaum [14] already observed that there are high frequences for conjunctive paraphrases for positive conditional threats ('if you do this I'll break your arm' becomes 'do this and I'll break your arm') (35%), positive conditional promises ('if you do this you'll get a chocolate' becomes 'do this and I'll get you a chocolate') (40%) and negative conditional promises ('if you don't cry I'll get you an icecream' becomes 'don't cry and I'll get you an icecream') (50%). However, he did not observe conjunctive paraphrases for contingent universals (where there is no intrinsic connection between antecedent and consequent) or even lawlike universals. Clearly, the statements we provided are contingent universals, so Fillenbaum's observations on promises and threats are of no direct relevance. However, if the course of event conditional is a possible reading of the conditional, the inference to a conjunction observed in many of our subjects makes much more sense. Clearly the truth conditions for conditionals of this type differ from the intended interpretation; to mention but one difficult case, when is a generic false? Thus, a generic interpretation may lead to different evaluations and selections. Here's an example of what this means in practice.

Example Subject 22 [experiment 2; subject has chosen the conjunctive reading in the booklet.]

> *E.* [Asks subject to turn the 5]
> *S.* That one ... that isn't true. There isn't an E on the front and a 2 on the back. [...] you turn over those two [E and 2] to see if they satisfy it, because you already know that those two [G and 5] don't satisfy the statement.
> *E.* [baffled] Sorry, which two don't satisfy the rule?
> *S.* These two don't [G and 5], because on one side there is G and that should have been E, and that [5] wouldn't have a 2, and that wouldn't satisfy the statement.

E. Yes, so what does that mean ... you didn't turn it because you thought that it will not satisfy?
S. Yes.

This provides a very interesting explanation for the p, q choice; these two are the only cards that could possibly satisfy the rule, $\neg p$ and $\neg q$ don't satisfy in any case, and so aren't to be turned according to the instructions! In our experiments, at least seven subjects who had selected p, q followed this line of reasoning. Several of these subjects said that they took the rule to be true, so they operate with a notion of truth that allows exceptions.

Example Subject 22 [continuation of previous quote; subject has said in the beginning 'I thought of that as a true statement']

E. And this one doesn't satisfy [G]?
S. No, because it's not an E.
E. But you still took the statement to be true.
S. Yes ... well my immediate reaction first time was to assume that this is a true statement, therefore you only turn over the card that you think will satisfy the statement.

A note of caution: one might be tempted to think that deontic conditionals also contain a strong generic element, so that there should be some analogy to the generic interpretation of indicative conditionals just outlined. But although in both cases cards can only obey or violate, the pattern of selections show that subjects still interpret these conditionals differently. Finally, even the case studied here, 'truth with exceptions' can mean different things to different people. Two subjects argued along the following lines:

Example Subject 18 [experiment 2]

S. If I just looked at that one on its own [5] I would say that it didn't fit the rule, and that I'd have to turn that one [E] over, and if that was different [i.e. if there wasn't a 2] then I would say the rule didn't hold.
E. So say you looked at the 5 and you turned it over and you found an E, then?
S. I would have to turn the other cards over ... well it could be just an exception to the rule so I would have to turn over the E.

We hope that at this point it hardly needs emphasising anymore that subjects may make the same selection of cards (e.g. p, q) for vastly different reasons, and that these reasons may be of more interest than the selections themselves.

6.5 The negation of a conditional

We also asked our subjects to determine what follows from the negation of a conditional: 'it is not the case that if there is a vowel on one side, then there is an even number on the other side'. Again, more than half of the subjects ticked the answer 'if there is a vowel on one side, there is an odd number on the other side'. We will refer to this as *strong negation*. This is in line with Fillenbaum's findings: he observes that in 60% of the cases the negation of a causal temporal conditional $p \to q$ ('if he goes to Amsterdam, he will get stoned') is taken to be $p \to$ not q; for contingent universals the proportion is 30%. In our experiment the latter proportion is even higher. Here is an example of a subject using strong negation when asked to imagine what could be on the other side.

Example Subject 26 [experiment 2; subject has chosen strong negation]

> *E.* So you're saying that if the statement is true, then the number [on the back of E] will be 2. ... What will happen if the statement were false?
> *S.* Then it would be a number other than 2.

Example Subject 18 [experiment 2; subject has chosen strong negation, has selected E and 2, thinks G is irrelevant]

> *E.* And the 5?
> *S.* It could have an E yes, but if that rule is true it will have another letter.
> *E.* And the 2?
> *S.* The 2 should have an E and if that rule is wrong it should have any other letter.

This finding may explain why some subjects think that turning only the p card suffices to establish truth or falsity in the standard task (in Wason's experiment, one-third of the subjects made this choice).[6] In our case however, although in the baseline task p was chosen as frequently as p, q, this response became rare after tutoring for the right interpretation

[6]It has sometimes been suggested (e.g. in Johnson-Laird and Byrne [23], p. 66) that strong negation is a consequence of taking the antecedent as a *presupposition*. This is analogous to Haiman's argument ([20], p. 583) that the antecedent of a conditional is a *topic* (in the technical sense): "A conditional clause is (perhaps only hypothetically) a part of the knowledge shared by the speaker and his listener. As such, it constitutes the framework for the following discourse". Apart from the notorious difficulties surrounding presupposition and topic, it seems to us that the dialogues suggest a different interpretation. Subjects apparently consider true and false to be symmetric; a false rule is one which is false of every instance.

of the anaphora, suggesting that the p response is due rather to constant anaphora.

The effect of strong negation on selection appears to be much more marked in condition 3, where at least four subjects say that *any* card can distinguish between the two rules. For instance, subject 22, who has strong negation, makes this choice. Observe that with strong negation the problem reduces to deciding between $p \leftrightarrow q, \neg p \leftrightarrow \neg q$ and $p \leftrightarrow \neg q, \neg p \leftrightarrow q$. Turning any card can indeed decide this.

7 Verifying and falsifying

This brings us to so called verification bias: subjects would be tempted to turn the q card, because q/p confirms the rule, whereas $q/\neg p$ is irrelevant. 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. Before we discuss this in detail, let us give an illustration

Example Subject 3 [experiment 1a] asked the question:

> *S.* Do I assume that I should turn the A over, since I know that on the back of the A is the 4. I have taken the rule to be true since it says that there is an A on one side and a 4 on the back. I guess that I should only turn over the ones that would potentially prove the rule.

Example Subject 13 [experiment 2]

> *S.* (Turns the 5) ... The card doesn't fit the rule.
> *E.* OK You didn't pick this card, the card you have just turned, are you still happy with your original choice?
> *S.* I thought I was trying to verify the statement rather than to falsify it. So you turned over the card that could falsify the statement, so no I suppose I'm happy with my first choice, although no, no, I was trying to verify with those letters, rather than falsify with those two ...[7]

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 establishing that the rule is true, then verification in this sense (i.e. seeking falsifying instances) is just what most subjects aren't doing, 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

[7]For another striking example, cf. subject 26 in section 9.

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 were seeking compliant cards, which cards are those? Clearly p/q cards are compliant. Clearly $p/\neg q$ cards are not compliant. In the transcripts, the vast majority of subjects regard both $\neg p/q$ and $\neg p/\neg q$ cards as neither compliant nor non-compliant but irrelevant. However, the trancripts also show that a sizable number of subjects make a distinction between $\neg p/q$, which is irrelevant, and $q/\neg p$, which falsifies! This shows, however, that turning the q card cannot always be considered as an instance of verification bias. In fact, about 40% of the subjects who chose p, q considered $q/\neg p$ to be falsifying. We shall now give some examples. The first example is of a subject denying verification bias.

Example Subject 1. [experiment 2]

> *E.* What could there be on the back of the 2?
> [Subject writes G and E.]
> *E.* OK. And in the case of the E?
> *S.* It wouldn't support it. Yeah it wouldn't support it. It wouldn't make it necessarily true.
> *E.* Let's consider them one at a time. What if there was an E on the back of the 2?
> *S.* Doesn't matter, doesn't either make it true or false.
> *E.* OK. So do you want to turn over the 2?
> *S.* Not particularly.

The next example presents again subject 4 who gave the normative response in experiment 1c, appears to be aware of the intended reading of anaphora and conditional and now struggles with experiment 2.

Example Subject 4.

> *E.* You picked the E, would you pick anything else?
> *S.* Yes I would [...] the G is irrelevant. OK the 5? [...] If there's a G there then that's fine, but if there's an E then that falsifies. [...] This one is a 2, it could be an E or there could be a G, so yes I would turn it over, if there was a G then that would falsify the rule. The G doesn't come into it because if there's a 2 - it doesn't say if there's a 2 there has to be a G. Does it ... shit ...
> *E.* Turn over the cards you want to turn.
> *S.* [picks up the 2] Well this falsifies the rule because ... no shit, does it?... Yes it does because there isn't an E and 2 combination.

E. Turn over the other cards now.
S. This is a 5. I have to turn it over to check whether there is an E. If there's an E then that also falsifies the rule. Oh and there is. I don't want to touch the G. Now I'm going to turn the E to see if there is a 2 on the other side, there is not.

The final example shows a case where the choice of the q card is an instance of looking for falsification, together with the explicit realisation that the conditional is uni-directional.

Example Subject 6.[experiment 1c]

S. [subject turns the A] Oh no. So that's wrong and that proves it wrong. ... [subject turns K] That doesn't really matter, does it? [goes over to the 4] I don't need to turn this? I do need to turn it. [turns over the 4, finding K] So that disproves it as well. [turns over the 7] So I could have picked this one [subject points to A].
E. Just that one?
S. Because they're the same [indicating A and 7] it must be wrong.
E. Well, it could have been different ...
S. It works that way [indicating with pen from left to right]. If there is a vowel on one side, then there is an even number on the other side, but if there is an even number on one side it doesn't necessarily mean that there is vowel on the other side.

This pattern of response is very common (≥ 40 %), and it casts a curious light on verification bias. Note that subject 4 selects the q card because it *potentially* falsifies; it is not just that a $q/\neg p$ result is *observed* to falsify the rule. The upshot seems to be that, contrary to what Wason believed, subjects selecting p, q do look for falsification, but they look for it in the wrong place. This cannot always be explained by assuming that subjects have a biconditional, because this is sometimes explicitly denied. Furthermore, they may evaluate the identical cards $\neg p/q$ and $q/\neg p$ differently. This has also been observed in the case of the $p/\neg q$ and $\neg q/p$ cards, but here it occurs less often (27%). As subject 4 shows, these asymmetries cannot always be explained by assuming that subjects fail to see the reversibility of the cards; she had the right understanding in experiment 1c and in experiment 2 she considered $p/\neg q$ and $\neg q/p$ to be equivalent.

We believe that it is this pattern of evaluations, rather than the pattern of actual card selections, that is one of the major riddles of the selection task. As indicated above, interference between anaphora and direction of implication might explain the observed pattern, but to substantiate this we would need independent evidence that subjects indeed decompose the anaphora while processing the task. This pattern also shows that explanations of good performance in thematic tasks using memory cueing miss the point: it is not that in (some) thematic (but not in abstract) tasks possible counterexamples can easily be retrieved from memory; rather, subjects consider different things to be counterexamples.

8 Social contracts and cheating detectors

So far, we have not considered how the various explanations explain whatever thematic effects have been observed, save perhaps for our oblique reference to the idea that some deontic conditionals tend to be interpreted biconditionally. However, as mentioned above, this tendency toward biconditional interpretation arises with closed-world readings, and most certainly won't explain the main observations of reasoning with deontic conditionals. The deontic conditionals (e.g. 'if the hiker stays overnight, he must bring firewood') are the rules where card selection is most normative. Social contract explanations focus almost entirely on thematic effects. Cosmides [10] original claim was that human beings, during their social evolution, developed 'cheating detector' algorithms which functioned to allow them to police social contract regulations (e.g. looking for a hiker staying overnight who has not brought his share of firewood), 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 would be the only mechanism with which undergraduate students (prior to logical instruction perhaps) can solve the task.

Cosmides rightly addresses the important issue of why it is that deontic material often leads to normative behaviour, but it seems to us that her proposed explanation cannot be upheld. Firstly, empirically, it is not true that only deontic material works. Several non-deontic contexts have been shown to facilitate normative reasoning, e.g. Sperber et al.[32] 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; and there is certainly something to be explained about the role of the deontic/indicative moods in these observations.

Secondly, however, the difference does not seem to lie in just cheating detection. As we have seen, even in the abstract rules under consideration, subjects who give the modal p, q response may motivate their choice by invoking falsification. So also in this case they look for 'cheaters', but in an odd place. It is this that has to be explained, not that only deontic material triggers cheating detection—because other material does too.

In fact the cheating detection hypothesis is much in need of clarification. How widely is it supposed to apply? Suppose the classical abstract task is run with a simple modification of the instructions which tell the subject that the source of the conditional rule is an inveterate (though unreliable) liar. Would one now expect subjects' cheating detectors to kick in and restore normative performance? If not, why not? Surely detecting liars is an important category of cheating detection? The social obligation toward truthfulness is foundational among social obligations. There has been remarkably little empirical curiosity shown toward the nature of 'evolutionary' explanations.

Thirdly, the use of deontic material also leads to a much-overlooked formal difference in the instructions. In the indicative case, one can ask the subject to

> select the cards you have to in order to determine whether the rule is true or false.

This makes no sense in the deontic case, where one can only ask to

> select the cards you have to turn in order to determine whether they obey the rule.

A critical difference between these two instructions is that the latter considers cards individually: a card obeys or disobeys independently of the other cards, whereas the first instruction implicitly requires an answer in terms of sets of cards. In particular, for those subjects who do not have strong negation, a single card can never conclusively falsify or verify. Some of these subjects, however, show clear signs of their struggle with dependencies between card selections:

Example Subject 10. [experiment 1c]

> *S.* OK so if there is a vowel on this side then there is an even number, so I can turn A to find out whether there is an even number on the other side or I can turn the 4 to see if there is a vowel on the other side.
> *E.* So would you turn over the other cards? Do you need to turn over the other cards?
> *S.* I think it just depends on what you find on the other side of the card. No I wouldn't turn them.
> ⋮
> *E.* If you found a K on the back of the 4?
> *S.* Then it would be false.

⋮

S. But if that doesn't disclude [*sic*] then I have to turn another one.
E. So you are inclined to turn this over [the A] because you wanted to check?
S. Yes, to see if there is an even number.
E. And you want to turn this over [the 4]?
S. Yes, to check if there is a vowel, but if I found an odd number [on the back of the A], then I don't need to turn this [the 4].
E. So you don't want to turn ...
S. Well, I'm confused again because I don't know what's on the back, I don't know if this one ...
E. We're only working hypothetically now.
S. Oh well, then only one of course, because if the rule applies to the whole thing then one would test it.

⋮

E. What about the 7?
S. Yes the 7 could have a vowel, then that would prove the whole thing wrong. So that's what I mean, do you turn one at a time or do you ...?

⋮

E. Well if you needed to know beforehand, without having turned these over, so you think to yourself I need to check whether the rule holds, so what cards do I need to turn over? You said you would turn over the A and the 4.
S. Yes, but if these are right, say if this [the A] has an even number and this has a vowel [the 4], then I might be wrong in saying "Oh it's fine", so this could have an odd number [the K] and this a vowel [the 7] so in that case I need to turn them all.
E. You'd turn all of them over? Just to be sure?
S. Yes.

Thus the necessarily different nature of the instructions adds a layer of complexity to the indicative case, as compared to deontic rules. Note also that deontic rules don't have explicit references to the cards, as do the abstract rules; and this could be of some importance, given that subjects struggle with the 'one side – other side' anaphor. In sum, it is not just deontic materials that work, and there are also formal differences between deontic and abstract tasks. There would seem to be little motivation to postulate evolutionarily advantageous 'cheating detectors', and little investment has been made in trying to flesh out just what characteristics such detectors might be expected to have. What is clear is that the nature of the various possible relations between cards and rule and between rule and cards is a rich source of these subjects' confusions.

9 Bayesianism and information value

We will now discuss in somewhat greater detail the Bayesian explanation of behaviour in the 4-card task due to Oaksford and Chater [29], because of its recent popularity. The point of departure of the Bayesian explanation 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 selection is also determined by factors different from logical evaluation. 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. One doesn't even have to call this a misinterpretation; as we have seen the course of events conditional actually invites a subject to consider a larger population.[8]

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 weighted by their probabilities. 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 (in this sense) 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

[8] This reading could also be suggested by an analysis of the conditional along the lines proposed by Lewis [26] which in our case would run as follows: '(Always: if x has a vowel one one side)(x has an even number on the other side)'.

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.

We shall now present an example which shows that subjects may indeed motivate their choices by pointing to what they perceive as the 'information value' of a card.

Example Subject 5. [experiment 2]

> E So you would pick the E and you would pick the 2. And lastly the 5?
> *S.* That's irrelevant.
> *E.* So why do you think it's irrelevant?
> *S.* Let me see again. Oh wait so that could be an E or a G again [writing the options for the back of 5 down], so if the 5 would have an E then that would prove me wrong. But if it would have a G then that wouldn't tell me anything.
> *E.* So?
> *S.* So these two [pointing to E and 2] give me more information, I think.
> *E.* [...] You can turn over those two [E and 2].
> *S.* [turns over the E]
> *E.* So what does that say?
> *S.* That it's wrong.
> *E.* And that one [2]?
> *S.* That it's wrong.
> *E.* Now turn over those two [G and 5].
> *S.* It's a G and 2. Doesn't say anything about this [pointing to the rule]. [After turning over the 5] Aha.
> *E.* So that says the rule is ...?
> *S.* That the rule is wrong. But I still wouldn't turn this over, still because I wouldn't know if it would give an E, it could give me an a G and that wouldn't tell me anything.
> *E.* But even though it could potentially give you an E on the back of it like this has.
> *S.* Yes, but that's just luck. I would have more chance with these two [referring to the E and the 2].

So in this case the evaluation of the $\neg q/p$ card is correct, but the selection differs from what is dictated by evaluation because the subject thinks that the chances of getting a counterexample with the $\neg q$ card are negligible. This is very interesting, because it lends some support to the analysis of the selection task in terms of information gain presented in Oaksford and Chater

[29].[9] Using a fair number of 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 as we have seen in at least some thematic 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, 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. Since we have concentrated on the abstract task here, we will not discuss utilties further.

We will now discuss the model in greater formal detail. Interestingly, it is adapted (cf. Oaksford and Chater [30]) 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. Similarly, subjects performing Wason's task would implicitly decide between the hypothesis of complete dependence (the foreground rule) and the hypothesis of independence.

In general, let X be an experiment with two outcomes, X_0 and X_1, designed to decide between hypotheses H_0 and H_1. Then one formula for the expected information gain upon performing X, $E_X(I)$, is given by

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

Let us now apply this formula to the 4-card task, pertaining to the implication $p \to q$. It is fundamental to Oaksford and Chater's [29] 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,

[9]Not unequivocally, however, because, as we have seen the q card may be selected for its potentially falsifying, not verifying, character.

this was not the way the task was specified in the instructions, but by thus misinterpreting the task, the subject naturally brings in probabilities and rival statistical hypotheses. Selecting a card and turning it over can be viewed as performing an experiment, which 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 $P(p),\ P(q)$[10], 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.[11]

Before we proceed to a methodological discussion, we give an example of a subject who explicitly weighs evidence *pro* and *con* in experiment 3, where the task is to decide which of two rules is true.

Example Subject 26 [experiment 3]

> *S.* [has turned U,I, found an 8 on the back of both] I can't tell which one is true.
> *E.* OK let's continue turning.
> *S.* [turns 3] OK that would verify rule two. [...] Well, there are two cards that verify rule two, and only one card so far that verifies rule one. Because if this [3] were verifying rule one, it should be an I on the other side.
> *E.* Let's turn [the 8].
> *S.* OK so that says that rule two is true as well, three of the cards verify rule two and only one verifies rule one.
> *E.* So you decide by majority.
> *S.* Yes, the majority suggests rule two.

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

[11] It is somewhat peculiar that Oaksford and Chater [30] refer to Horwich' *Probability and evidence* [22] for a fuller treatment of Mackie's solution of the ravens paradox, whereas Horwich is at pains to argue that Mackie's solution is wrong. In fact, arguing along Horwich' lines would lead to the conclusion that the $\neg q$ card is *more* informative than the q card.

It is highly interesting that 3/U is described as *verifying* rule two, rather than *falsifying* rule one; U$\rightarrow$8 is never ruled out:

> *S.* It's not completely false, because there is one card that verifies rule one.

Asked to describe her thought processes, the subject later comments

> *S.* Well when there's two rules then you can't say that they should both be true because they are mutually exclusive ... so depending on which way the cards are there is basically a 50% probability that either one is going to be true. [...] With one rule I think it will be true or if it wouldn't be true, then it seems more likely that it would be true.

We performed experiment 3 because we were interested how subjects would reason when they were explicitly presented with two rival hypotheses. Apart from subject 26, subjects tried to solve the task by logical processing. In fact, after having turned the cards, 7 out of 10 subjects concluded that only the 3 card is relevant (not surprisingly, subject 26 never reached this stage). It is not in the spirit of this paper to argue that 'Bayesian' processing doesn't happen, but we can say that it doesn't show itself in many subjects. Also in the standard task, it seems that whenever an alternative to $p \rightarrow q$ is considered, it is not 'p, q are independent', but rather $p \rightarrow \neg q$, as argued in section 6.5.

We will now proceed to give a brief methodological discussion of the Bayesian approach, to acquaint the reader with the kind of assumptions that have to be made in order to get the model to work.

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, as Laming [25] 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. Put another way, this kind of application of Bayesian theory inherently ignores individual differences in behaviour. Our dialogue evidence strongly suggests that these differences are not mere noise but rather are a significant part of what needs to be explained about human reasoning. Subjects make different interpretations and representations of this context and their different behaviour results.

To see the model at work, consider what the predictions are when the rule is varied by introducing negations in antecedent and/or consequent. This 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' ($\neg p \rightarrow q$). The observed 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 [29] offer two solutions here. The first derives from Oaksford and Stenning [28] 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 some 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 'p and q are independent', and H_1 is '$p \rightarrow q$', 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 $\neg p \rightarrow q$ forces the inequality $P(q) \geq P\neg(p)P(H_1')$, where H_1' says that $\neg p$ is contained in q. This boundary 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.

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. (Note also that the parameter values taken to characterise the environment are not empirically determined.)

In this respect it is of interest to discuss Oaksford and Chater's [30] reaction to an experiment of Pollard and Evans (for a discussion, see Evans and Over [13]), 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 $p \to q$) 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 usually 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.

To us, this suggests that, while there is evidence that some subjects engage in some *qualitative* form of Bayesian processing, it is useless to try to fit all observed behaviour into one *quantitative* model. The number of assumptions necessary to make the model work is so large that the model looses all explanatory power. Nor is this all.

The most telling 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 modulo a small set of exceptions, where in the latter 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. As an illustration, we consider van Duyne's experiments [38]. He compared four different formulations of a conditional statement in 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:

I overall, there would be a significant difference between abstract and thematic materials;

II 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;

III 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[12] , in the sense that the percentage of correct answers is the same.[13] 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 one 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 logically equivalent sentence types would be treated the same 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. In sum, the fundamental shortcoming of the Bayesian model is that it is by and large insensitive to meaning.

10 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. The terms in the instructions that may be subject to different interpretations include 'true', 'false', 'obey', 'fit', 'violate' and 'have to turn'. Furthermore, it is possible that some terminology is more conducive to good performance than other. Lastly, under this heading we also include factors which might be more appropriately called pragmatic, namely the interaction between experimenter and subject. Specifically, one may think of the difference between cooperative and adversarial communication: the subject is asked to take the background rule on trust, but must be agnostic about the foreground rule. A highly authoritative experimenter figure (who is quite correctly taken by the subjects to be omniscient with regard to the materials) states two rules, but indirectly indicates that one of them is not trustworthy. This is itself a strange communicative act—"I'm

[12]In our own paraphrase experiment, subjects showed themselves wary of disjunctions. When the target sentence involved a disjunction, few subjects selected correct paraphrases involving other logical constants; and when a correct paraphrase was formulated in terms of a disjunction, most subjects failed to select it. This suggests that $p \to q$ is not perceived as being equivalent to $\neg p \vee q$ and furthermore that it is hard to process the latter form.

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

going to tell you two things. Trust one, but the other might be a lie." The pragmatics of this aspect of the deontic cases is quite different. Here it is the miscreants in the pub who might be breaking the law, not the experimenter who might be purveying a falsehood. Again one wonders how performance would be affected if the source of the foreground rule were different than the experimenter; the source were known to be unreliable; or even if the experimenter were clearly ignorant of the materials.

10.1 Relations between cards and rules

A key issue in 4-card task performance appears to be the possible differences in perspective on the semantic relations between rule and card, or between 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. This is not merely a difference in perspective that makes no difference to outcome, and it is one way in which deontic and indicative rules differ. In the case of a deontic rule, it makes no sense to inquire about its truth or falsity, one may only ask whether cases obey or violate the rule. Thus, one views the cards from the perspective of the rule. For indicative rules, truth may be a real issue, and card-rule and rule-card relations are only sometimes symmetrical. The rule can make the cards into counterexamples, but the rule can also be viewed from the perspective of the cards, which can make it false. With compliant cases, the rule can make the case fit. But the case alone cannot make the rule true by fitting—only sets of cases can do that. As one subject put it:

Example Subject 16 [experiment 1c]

> *E.* If there was a 7 on the back of the A, what would that mean?
> *S.* It would mean that the rule is false.
> *E.* Would it mean that this card doesn't fit the rule or that the rule is false?
> *S.* They are both different sides of the same coin. Mind you, it would suggest that the rule was wrong more than the card was wrong. The cards would be what the rule would be drawn from.

Here is an example of the opposite perspective

Example Subject 8 [experiment 2; has turned the 5 to find E]

> *S.* Turn this over and the rule is wrong.
> *E.* The question is different, whether the cards fit the rule or not.
> *S.* Sorry yes, this is an autonomous rule which can occur for some cards. But these cards don't fit the rule.

Now as we indicated above, in section 8, the perspective from the rule appears to be simpler than the perspective from the cards. In the first case, one is at liberty to consider the cards independently. In the second case, issues of dependency arise, and subjects sometimes struggle with the instruction to select the cards they *have to* turn; after all, if turning the p card shows the rule to be false, do you have to turn the $\neg q$ card as well?

One possibility to improve performance on the indicative task would thus be to induce subjects to take the perspective from the rule. A variant of this idea did enjoy an outing in the literature, starting out from Yachanin and Tweeney [45], 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; or with a 'reduced array selection task' which presented only consequent-visible cards (q and $\neg q$); or with the extra task of providing verbal justifications, did it increase the turning of $\neg q$. Hence this way of changing the perspective was not effective.

Neither did our manipulation of asking subjects 'to select the cards they have to turn to determine whether they *obey* the [indicative] rule' succeed in improving performance. What became clear though is that subjects in this condition tend to adopt a curious generic reading of the conditional, as explained in section 6.4, and that their choices are consistent with this reading. The rule is allowed to have exceptions, but due to the conjunctive reading, $\neg q$ doesn't obey in any case, hence is not chosen (e.g. this is true for subject 8, who has the correct perspective). Hence even when the instructions manage to induce a generic reading, it is different from the one adopted in the deontic case.

10.2 Cooperative and adversarial communication

One source of confusion in the task is, surprisingly, that subjects are unclear about the status of the foreground rule: must they take it as true (on a par with the background rule) or is its truth value in doubt? Even when subjects are explicitly instructed to determine the truth value of the rule, they may still take it to be true. For example

Example Subject 3 [experiment 2; has chosen E and 2]

> *E.* Why pick those cards and not the other cards?
> *S.* Because they are mentioned in the rule and I am assuming that the rule is true.

One advantage of the task where subjects have to determine which of two rules is true, and which false (experiment 3), is that this confusion can no longer arise. The transcripts show that this leads to an increase in 'logical' processing during tutoring. (Only subject 26 engaged in nonlogical processing, as illustrated in section 9.) Subjects select cards based on their evaluations; in this case these evaluations are often wrong because they implicitly adopt strong negation, or in other words because they have the wrong concept of what it means for a rule to be false. This misconception is exposed when they have turned both the U and the I to find an 8; 7 out of 10 subjects then spontaneously realise that they only have to turn the 3. But matching bias does not occur here, nor is there any evidence of a conflict between evaluation and selection. Hence clarifying the intention by using a task in which this aspect of the instructions is clear can lead to an improved performance, even in the indicative case.

11 Conclusions

The exploratory experiment reported here has cast its nets somewhat wider than is customary, to obtain information about subjects' processing and semantic understanding of the task. The picture that emerges is complex. The differences between subjects, even when they make the same selection, are huge and defy any single explanation. In fact, there appears to be no explanation for a very common pattern of evaluation and selection: p, q is selected, $q/\neg p$ is evaluated as falsifying and $\neg p/q$ is evaluated as irrelevant, although we ventured a hypothesis in section 6.3. It does not seem very helpful to describe such behaviour as 'irrational', it is more interesting to relate this and other behaviour to subjects' understanding of the task. We have seen that this strategy sometimes leads to clarification, as when subjects' understanding of 'false' helps to explain their performance in experiment 3. It would be very useful to do the same series of experiments again with thematic, in particular deontic, material, to isolate the stages where processing begins to differ from what has been observed in the abstract task.

Our investigations may be seen as falling under the general heading of 'rational analysis' (cf. Anderson [2], Oaksford and Chater [29]), i.e. the goal of understanding how subjects assimilate the tasks set to them in the settings that prevail through understanding their relations to other contexts of communication. Where we differ from the authors mentioned is that we believe the aim requires investigation of a wider range of settings, using a broader range of data, and also requires faithfulness to a greater range of analyses of language and communication. There is a danger that deceptively simple statistical models obscure the phenomena in need of explanation, and that seeing subjects' assimilation of the task to general information seeking patterns dissmisses the educational relevance of the logical competence models and their highly objectified stance toward language. Stanovich and West [34] show how closely related this stance is to other educational achievements. The tutorial dialogues presented here provide some insight into the variety of students' problems which may be of some help to those involved in teaching these skills.

Furthermore, these dialogues may also provide challenges to natural language semantics. While it is of course possible to attribute the vacillations in interpretation (of conditionals and anaphora, for example) to performance factors, it seems more interesting to look into the structure of the linguistic competence model to see how the observed interferences may arise. It seems to us that dialogues such as these provide a rich source of data for semantics and pragmatics, which promises to yield deeper insight into interpretation and processing of natural language.

What we hope to have demonstrated in this preliminary study is that the data do not warrant abandoning the search for formal models to provide bases for explaining subjects' reasoning behaviour. Instead, formal models embodying insights from neighbouring fields are useful guides for a richer program of empirical exploration and testing.

12 Appendix A: Classroom Experiment Data

Table 1 displays the data from the initial classroom experiment comparing selections for four different rules as specified in 6.1.

Response	Instruction				
	classical	reversible	constant-front	constant-back	Total
0000	3	2	3	0	8
0010	2	0	1	3	6
0100	0	0	1	0	1
0101	0	2	0	0	2
0110	0	0	0	1	1
1000	5	4	1	3	13
1001	1	2	3	1	7
1010	9	8	9	8	34
1011	0	1	2	0	3
1100	1	2	1	0	4
1111	0	0	0	2	2
Total	21	21	21	18	81

Table 1: Frequencies of responses for the four different rule in the conditions 6.1. '0' in the response label indicates no turn; '1' indicates subject turns card (cards in the usual $p, \neg p, q, \neg q$ order).

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] A. Athanasiadou and R. Dirven. Typology of *if*-clauses. In E. Casad, editor, *Cognitive linguistics in the redwoods*, pages 609–654. Mouton De Gruyter, 1995.

[4] A. Athanasiadou and R. Dirven. Conditionality, hypotheticality, counterfactuality. In A. Athanasiadou and R. Dirven, editors, *On conditionals again*, pages 61–96. John Benjamins, 1997.

[5] A. Athanasiadou and R. Dirven. *On conditionals again.* John Benjamins, Amsterdam, 1997.

[6] 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.

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

[8] COBUILD. *Collins Birmingham University International Language Database.* Collins, 1980.

[9] B. Comrie. Conditionals: a typology. In E. Traugott, A. ter Meulen, J.S. Reilly, and C.A. Ferguson, editors, *On conditionals*, pages 77–99. Cambridge University Press, 1986.

[10] 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.

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

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

[13] 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.

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

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

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

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

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

[19] 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.

[20] J. Haiman. Conditionals are topics. *Language*, 54:564–589, 1978.

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

[22] P. Horwich. *Probability and evidence.* Cambridge University Press, Cambridge, 1982.

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

[24] P.N. Johnson-Laird and P.C. Wason. A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1:134–148, 1970.

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

[26] D. Lewis Adverbs of quantification in E. Keenan (ed.), *Formal semantics of natural language* Cambridge University Press, 1975.

[27] 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.

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

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

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

[31] N.S. Smalley. Evaluating a rule against possible instances. *British Journal of Psychology*, 165(2):293–304, 1974.

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

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

[34] K. E. Stanovich and R. F. West. Cognitive ability and variation in the selection task. *Thinking and reasoning*, 4: 193–230, 1998.

[35] 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.

[36] K. Stenning and J. Levy. Knowledge-rich solutions to the binding problem: a simulation of some human computational mechanisms. *Knowledge Based Systems*, 1(3), 143–152, 1988.

[37] E. Traugott, A. ter Meulen, J.S. Reilly, and C.A. Ferguson. *On conditionals.* Cambridge University Press, Cambridge, 1982.

[38] P. C. van Duyne. Realism and linguistic complexity in reasoning. *British Journal of Psychology*, 65(1):59–67, 1974.

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

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

[41] P. C. Wason and D. W. Green. Reasoning and mental representation. *Quarterly Journal of Experimental Psychology*, 36A:598–611, 1984.

[42] 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.

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

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

[45] S.A. Yachanin and R.D. Tweeney. The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society*, 19:87–90, 1982.