

Semantics and cognition: semantic insights into the  
psychology of reasoning  
DRAFT OF BOOK: DO NOT QUOTE!

Keith Stenning      Michiel van Lambalgen

June 21, 2002



# Contents

<b>1</b>	<b>Logic and psychology</b>	<b>5</b>
1.1	Logic in daily life . . . . .	6
1.2	What is logic? . . . . .	8
1.3	(Anti-)Psychologism . . . . .	8
1.3.1	Frege . . . . .	9
1.3.2	Husserl . . . . .	9
1.4	Aims of the psychology of reasoning . . . . .	11
1.4.1	Some competing schools . . . . .	11
1.5	A revised view . . . . .	12
1.6	Antropological data . . . . .	13
<b>2</b>	<b>Syllogisms and beyond</b>	<b>15</b>
2.0.1	Euler's circles . . . . .	17
2.1	Individual differences in reasoning . . . . .	24
2.1.1	Immediate inferences and the understanding of quantifiers . . . . .	24
2.1.2	Syllogistic reasoning . . . . .	26
2.2	'Mental models' . . . . .	31
2.2.1	Full predicate logic . . . . .	40
<b>3</b>	<b>Propositional logic—the easy cases</b>	<b>45</b>
3.1	Inference tasks . . . . .	45
3.1.1	Children's propositional reasoning . . . . .	48
3.2	'Suppression effect' and nonmonotonicity . . . . .	49
3.3	'illusory inferences' . . . . .	63
<b>4</b>	<b>Propositional logic—the hard cases</b>	<b>67</b>
4.1	Wason's 4 card task . . . . .	67
4.1.1	Deontic variants . . . . .	68

4.1.2	Tutorial experiments . . . . .	71
4.1.3	Standard explanations, and why they fail . . . . .	73
4.1.4	Verifying and falsifying . . . . .	74
4.2	2-rule selection task . . . . .	77
4.3	Subjects' understanding of truth and falsity . . . . .	78
4.3.1	The logic of 'true' . . . . .	78
4.3.2	Dependencies between card-choices . . . . .	86
4.3.3	The pragmatics of the descriptive selection task. . . . .	90
4.3.4	Subjects' understanding of propositional connectives . . . . .	91
4.4	Experiment . . . . .	93
4.4.1	The Conditions . . . . .	93
4.4.2	Subjects . . . . .	97
4.4.3	Method . . . . .	97
4.4.4	Results . . . . .	97
4.5	The meaning of conditionals . . . . .	99
4.6	Anaphora . . . . .	101
4.6.1	Appendix A: Classroom Experiment Data . . . . .	112
<b>5</b>	<b>Logic and probability</b>	<b>113</b>
5.1	Probability underlies logic . . . . .	113
<b>6</b>	<b>Logic and evolution</b>	<b>123</b>
6.1	Evolutionary thinking: from biology to evolutionary psychology.123	
6.1.1	Adaptations and exaptations . . . . .	124
6.1.2	Massive modularity . . . . .	126
6.1.3	No role for logic? . . . . .	126
6.1.4	Cheater detection . . . . .	127
6.2	Logic and evolution—experimental data . . . . .	128
6.2.1	Social contracts and cheating detection . . . . .	128
<b>7</b>	<b>Logic and the brain</b>	<b>137</b>
7.1	Avenues to brain-correlates of reasoning . . . . .	138
7.1.1	Positron emission tomography (PET) . . . . .	142
7.1.2	Functional magnetic resonance imaging (fMRI) . . . . .	142
7.2	Experimental results . . . . .	142
7.2.1	Unilateral brain lesions and the Wason selection task . . . . .	142
7.2.2	ECT and syllogistic reasoning . . . . .	143
7.2.3	Brain-imaging studies of deduction . . . . .	144
<b>8</b>	<b>Autism, development and evolution</b>	<b>147</b>

# Chapter 1

## Logic and psychology

The purpose of this course is twofold. Our first aim is to see to what extent the psychology of reasoning and logic (more generally, semantics) are relevant to each other. After all, the psychology of reasoning and logic are in a sense about the same subject, even though in the past century a rift has opened up between them. Very superficially speaking, logic appears to be normative, whereas the psychology of reasoning is descriptive and concerned with processing. The first question then is: what is the relation between these two fields of inquiry? A wider aim is to discuss some of the theories offered in the literature from the point of view of the philosophy of science. Both mental models theory and evolutionary psychology, which takes its starting point in an observation about the psychology of reasoning, have become hugely popular explanatory paradigms in psychology. We will see that the experiments claimed to support these theories are marred by grave methodological errors, and that the theories themselves show little awareness of the subtleties of logic. So why their popularity? This will lead us to more general considerations about the roles of empirical evidence, concept formation and ideology in science.

Traditionally, it has been assumed that the results obtained in the psychology of reasoning tell us something about the rationality, or rather the absence thereof, of human reasoning. The following extended quotation from one of the founding fathers of the field, Peter Wason, exemplifies this attitude to perfection. He writes, concluding an overview of his ‘selection task’ paradigm for *The Oxford Companion to the Mind*

Our basic paradigm has the enormous advantage of being artificial and novel; in these studies we are not interested in everyday thought, but in the kind of thinking which occurs when

there is minimal meaning in the things around us. On a much smaller scale, what do our students' remarks remind us of in real life? They are like saying 'Of course, the earth is flat', 'Of course, we are descended from Adam and Eve', 'Of course, space has nothing to do with time'. The old ways of seeing things now look like absurd prejudices, but our highly intelligent student volunteers display analogous miniature prejudices when their premature conclusions are challenged by the facts. As Kuhn has shown, old paradigms do not die in the face of a few counterexamples. In the same way, our volunteers do not often accommodate their thought to new observations, even those governed by logical necessity, in a deceptive problem situation. They will frequently deny the facts, or contradict themselves, rather than shift their frame of reference.

Other treatments and interpretations of problem solving could have been cited. For instance, most problems studied by psychologists create a sense of perplexity rather than a specious answer. But the present interpretation, in terms of the development of dogma and its resistance to truth, reveals the interest and excitement generated by research in this area. (Wason [111, p. 644])

This is not the way in which we will approach the topic here. We believe that the outcome of the reasoning process is determined more by semantic, pragmatic, and processing factors than by the factors brought up by Wason, which belong to the domain of social psychology. Wason's idea was that 'minimising meaning' would allow us to discover reasoning in its pure form. By contrast, we believe, and claim to have shown, that minimising meaning leads to a furious interplay between all the factors mentioned. The main effect of this interplay is that the notion of logical form becomes problematic. One cannot simply read off the logical form from a given sentence, but the logical form is imposed on the sentence under the influence of semantics, pragmatics and processing limitations. Logical form will accordingly be our guiding theme in these notes.

## 1.1 Logic in daily life

The psychology of reasoning is not (primarily) concerned with formal reasoning as it occurs in, say, mathematics, but with reasoning in daily life. There, logic often underlies other reasoning activities (not necessarily de-

ductive), such as abduction (e.g. reasoning from effect to cause), or natural language interpretation. An example of the latter (discussed by Hans Kamp in a lecture):

A father and his son were sitting in a pick-up truck. The younger one was wearing a hat.

We effortlessly identify the younger one with the son. Let's see what logical reasoning goes on 'behind the scenes'. General knowledge gives:

$$\forall x(\text{wears a hat}(x) \rightarrow \text{human}(x)).$$

It follows by restriction to the domain of discourse that

$$\forall x(\text{wears a hat}(x) \rightarrow \text{father}(x) \vee \exists y(\text{father}(y) \wedge \text{son}(x, y)))$$

If we define *the younger one* by

$$\text{the younger one}(x) \leftrightarrow \neg \exists y \text{younger}(y, x),$$

the second premise gives

$$\forall x(\text{wears a hat}(x) \rightarrow \neg \exists y \text{younger}(y, x),$$

and since

$$\forall x \forall y(\text{son}(x, y) \rightarrow \text{younger}(x, y)),$$

it follows that

$$\forall x(\text{wears a hat}(x) \rightarrow \exists y(\text{father}(y) \wedge \text{son}(x, y))).$$

Here is an example of a complicated 'deductive' argument from ordinary (legal) life, which is almost a parody of logic (see Skalak and Rissland [86]). Under the heading 'Recognized forms of legal argument: a partial inventory', they include the '*Turkey, chicken and fish*' or '*double negative*' argument:

This convoluted argument takes the following form: the case at bar is so *unlike* the cases where a rule's conditions were held *not* to have been established that the rule *should* apply to the current case. The argument may be used more generally to claim that an instance is within some category, whether or not that category is implicitly defined by the rule consequent. This slightly expanded form of the argument maintains that something is so unlike negative examples of the category—unlike things outside

the category—that it should be considered a positive example—as within the category.

The name ‘turkey, chicken and fish’ stems from a hypothetical example in which a turkey farmer in Delaware tries to receive an entitlement reserved by the Federal Department of Agriculture regulation for chicken farmers. If the only cases denying the entitlement dealt with fish farmers, then the turkey farmer could argue that his turkey farms are so unlike fish farms that the FDA regulations should apply to his turkey business as well.

## 1.2 What is logic?

Logicians are familiar with a great many formal systems: classical predicate logic, intuitionistic logic, relevance logic, deontic logic, . . . . Accordingly, there is hardly a temptation anymore to identify ‘logic’ with a particular formal system. It is more common nowadays either to define a logical system as a consequence relation  $\Delta \vdash \Gamma$  satisfying certain ‘primitive’ conditions on *vdash* (cf. Gabbay [30]), or to view a logic as a certain relation between formulas and models; cf. Barwise and Feferman [7] for this point of view.

Needless to say, this is not the view of logic adopted by the psychology of reasoning. On the contrary, it has adopted a normative stance: there is one right logic (classical logic) and failure to comply with its norms lays one open to the charge of irrationality. Given the ubiquity of deviations from the norm, the associated research program is therefore concerned with finding explanations for the deviations observed. Thus, the psychology of reasoning is proceeding as if nothing has happened in logic in the past century. In part logic itself is to blame for this: it turned its back on psychology as soon as it could<sup>1</sup>

## 1.3 (Anti-)Psychologism

**Psychologism** In its naturalistic guise, this doctrine holds that thinking and knowledge are mental/psychological phenomena and that therefore logical laws are psychological laws. E.g.  $\neg(A \wedge \neg A)$  represents the impossibility of thinking contradictory thoughts at the same time.

---

<sup>1</sup>For an account of some of the factors which made psychology turn away from logic, see Stenning and van Lambalgen [100].

### 1.3.1 Frege

Frege had two main reservations about a naturalistic treatment of logic (and mathematics): (1) psychologism makes logic pertain to ideas only, so that it might seem that its relevance to the real world is nil

Psychological treatments of logic . . . lead then necessarily to psychological idealism. Since all knowledge is judgmental, every bridge to the objective is now broken off. (G. Frege, *Nachgelassene Schriften*; see [29])

The logicians . . . are too much caught up in psychology . . . Logic is in no way a part of psychology. The Pythagorean theorem expresses the same thought for all men, while each person has its own representations, feelings and resolutions that are different from those of every other person. Thoughts are not psychic structures, and thinking is not an inner producing and forming, but an apprehension of thoughts which are already objectively given. (G. Frege, letter to Husserl; see Vol.VI, p. 113 of [53])

(2) logical and mathematical knowledge is objective, and this objectivity cannot be safeguarded if logical laws are properties of individual minds

Neither logic nor mathematics has the task of investigating minds and the contents of consciousness whose bearer is an individual person. (G. Frege, *Kleine Schriften*; see [29])

If we could grasp nothing but what is in ourselves, then a [genuine] conflict of opinions, [as well as] a reciprocity of understanding, would be impossible, since there would be no common ground, and no idea in the psychological sense can be such a ground. There would be no logic that can be appealed to as an arbiter in the conflict of opinions. (G. Frege, *Grundgesetze der Arithmetik*; see [29])

### 1.3.2 Husserl

A strikingly modern view of logic is to be found in the writings of Husserl, in particular the *Logische Untersuchungen*, [52]. He views logic as theoretical discipline concerned with ‘truth’, ‘judgement’ and similar concepts. He grounds the normative status of logic via a combination of the theoretical statement ‘only such and such arguments preserve truth’ and the normative

statement ‘truth is good’. The so-called ‘Rules of logic’ are therefore subservient to description of logical concepts (such as truth). This immediately implies a criticism of psychologism: logical laws are exact and unassailable, empirical laws are approximative and provisional. And what *is* the empirical status of logical laws? How would they be learned? Psychologism leads to relativism: every human being has its own truth<sup>2</sup>—however, truth is absolute. In particular, psychologism claims its own truth, and is therefore selfrefuting.

The above naturalistic version of psychologism should be distinguished from ‘transcendental psychologism’. This doctrine, which can be found (in different forms) in Kant and Husserl, maintains that there are universal mental laws which hold for each subject *in virtue of being a subject*. This is why intuitionism is not subject to critiques of psychologism—mathematical constructions are universal because they depend on universal properties of the subject.

**Comments** Frege and Husserl wrote at a time before metalogic was developed. That is, although they knew a logical system corresponding to full predicate logic, they did not conceive of this system in a semantic manner. We now think of a valid formula is a formula true in all models, but this would not have made sense to Frege and Husserl, who could not conceive of a plurality of models. The metalogical perspective, however, has drastically altered our view of the normative character of logic. We conceive of a plurality of models, a plurality of semantic interpretations, and a plurality of definitions of validity. Each choice among these (if consistent) determines a different logic, in the sense of a set of valid arguments. Glib talk of the ‘apodictic certainty of logic’ is therefore out of the question. It is indeed a matter of *mathematical* certainty that once some parameters are fixed (choice of model-type, semantic interpretation, definition of validity), the logic is completely determined. But the parameters have to be set—by the discourse, or the social context. Setting the parameters is itself a process which requires reasoning. Thus the modern view of logic leads automatically to a duality between reasoning *for* an interpretation and reasoning *from* an interpretation.

**The importance of strategy** Traditionally, the psychology has focussed on the *set* of valid arguments associated to a particular logic. However, it

---

<sup>2</sup>Sometimes this is cheerfully accepted: the logician Dov Gabbay said in an interview: ‘everybody his own logic!’.

should have concerned itself with the (or rather *a*) ‘theorem prover’ giving that set. [to be amplified]

## 1.4 Aims of the psychology of reasoning

The psychology of reasoning is customarily viewed on a par with fields such as psychology of vision: tries to identify a *mechanism*. Psychologists typically ask: what is the fundamental reasoning mechanism? What do people *do* when solving a reasoning task? In which terms should an answer be given? Before we consider whether this question is necessarily the one most appropriate, let us consider several proposed answers.

### 1.4.1 Some competing schools

**Mental logic** The important names here are the recently deceased M.D.S. Braine [10] and L. Rips [81], [82]. This school maintains that logical reasoning is the application of formal rules, more or less like natural deduction. Here is an example: the theory tries to explain why humans tend to have difficulty with *modus tollens*, by assuming that this is not a primitive rule, unlike *modus ponens*; *modus tollens* has to be derived each time it is used, therefore it leads to longer processing time.

**Mental models** The founding father of the ‘mental models’ school is P.N. Johnson-Laird; cf. his book [55]. The main claim of this school is that reasoners do not apply content-independent formal rules (such as e.g. *modus ponens*), but construct models for sentences and read off conclusions from these, which are then subject to a process of validation by looking for alternative models. Errors in reasoning are typically explained by assuming that subjects read off a conclusion from the initial model which is not true in all models of the premises. The ‘mental models’ school arose as a reaction against ‘mental logic’ because it was felt that formal, content-less rules would be unable to explain the so-called ‘content-effects’ in reasoning; these will be explained in section 4.1.

**Darwinian algorithms** Evolutionary psychology has also shed its light on logic, beginning with the famous (or notorious) paper ‘The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task’ by Leda Cosmides [19]. Here the main claim is that there is no role for (formal, domain-independent) logic in cognition;

whenever we appear to reason logically, this is because we have evolved strategies ('Darwinian algorithms') to solve a problem in a particular domain (such as social contracts).

There are more 'schools' than have been mentioned here. Some will be treated below, for example the Bayesian approach (see section 5.1), which holds that what seems like logical reasoning really is probabilistic reasoning. Limitations of time and space have prevented us from discussing the pragmatic reasoning schemas of Cheng and Holyoak [15]. The views discussed were selected with the aim of illustrating the tension between form and content that pervades psychological thinking about reasoning. 'Mental logic' and 'Darwinian algorithms' are at opposite ends of the scale here, and 'mental models' professes to be somewhere in the middle<sup>3</sup>.

## 1.5 A revised view

As will gradually become clear, the point of view advocated in these notes differs from all these 'schools'. In a nutshell, we believe that the importance of semantics has been vastly underrated by the various theories. This is connected to another point: the relationship between reasoning and interpretation. The standard view of that relationship is, that the semantic interpretation of the material in a reasoning task precedes reasoning with that material. This has by and large been taken as an excuse for not bothering about semantics when studying reasoning. To be sure, when studying conditional reasoning the possibility has been entertained that in some cases a conditional is read biconditionally; but usually both are read materially, and the many careful discussions of conditional meaning to be found in the linguistic literature are not taken into account at all.

If one lets go of the assumption that there is only one possibility for interpretation, then solving a reasoning task also involves the imposition of a logical form on the task, a process which itself involves (meta-)reasoning. ('The experimenter must mean that ...') Thus, there is both reasoning *for* an interpretation and reasoning *from* an interpretation. Note that logical form involves: at least the syntactic form of statements, and the relation between syntax and models (i.e. a truth definition, including the specification of logical constants).

When experimentally investigating reasoning behaviour, a whole complex of phenomena is therefore investigated simultaneously: the imposi-

---

<sup>3</sup>Although one outcome of the discussion will be that it is hardly distinguishable from 'mental logic'.

tion of logical form, the mental representation of that logical form, memory-limitations interfering when doing this, drawing consequences by means of a theorem prover, ... The problem is: how to tease these apart?

## 1.6 Antropological data

Some information about the various factors that go into solving a reasoning task, can be obtained from anthropological studies. The paradigmatic study here is Luria's research among literate and illiterate populations in Kirghizia and Uzbekistan in the 1930's (see [69]). He showed subjects an argument such as the following:

In the Far North, where there is snow, all bears are white.  
Novaya Zemlya is in the Far North, and there is always snow  
there.  
What color are the bears there?

Literate subjects had no trouble answering this question, but illiterate subjects could come up with answers such as: 'How should I know what color the bear was? I haven't been to The North.'

The same phenomena were observed by Scribner in her work among the illiterate Kpelle tribe in Liberia (see scri97:mind). Here is a sample argument

All Kpelle men are rice farmers.  
Mr. Smith<sup>4</sup> is not a rice farmer.  
Is Mr. Smith a Kpelle man?

Again, subjects refused to answer the question definitively, instead giving evasive answers such as 'If one knows a person, one can answer questions about him, but if one doesn't know that person, it is difficult.' Scribner then went on to show that a few years of schooling in general led to the competence answer.

Apparently, the illiterate subjects do not understand what is being asked of them: to answer the question solely on the basis of the premises given. It is hard to blame them. After all, if the subject takes the questioner seriously, he wants to give a truthful answer (one of Luria's subjects said: 'I don't want to lie'). Even if the subject would understand the logical game, the argument would reduce the truth of the conclusion to that of the premises, for which the subject claims not to have evidence. The communicative situation

---

<sup>4</sup>'Mr. Smith' is not a possible Kpelle name.

is a bit strange: the subject is required to answer a question truthfully on the sole basis of information supplied by the questioner, abstracting from his own knowledge. Indeed, this is precisely the situation during an examination, so it is small wonder that a few years of schooling leads to improved performance on such tasks.

This little excursion into anthropology exposes some of the pitfalls of the experimental psychology of reasoning. For instance, Scribner argued that the results outlined above do not indicate defective logical skills, but rather point to a tendency not to apply logical skills to familiar material, of which the subject has knowledge beyond that stated in the premises. If that knowledge does not suffice to settle the issue, the subject will refrain from giving an answer. By contrast, subjects would have no trouble reasoning from completely new material, where they do not have to opportunity to consult their own knowledge. This hypothesis was corroborated in a study by Tulviste [104], who showed that Nganasan<sup>5</sup> children attending primary school failed on logical reasoning tasks involving daily activities, but did well on tasks involving typical school material (his example featured the metal molybdenum, with which the children would have little acquaintance). This observation is especially interesting in the light of claims that good performance on the notoriously difficult Wason selection task (see section 4.1) is *facilitated* by familiar material. Thus, it would be easier to focus on violations of the rule

If you want to drink alcohol, you have to be over 18

than to come up with counterexamples to an abstract rule such as

If there is a vowel on one side of the card, the other side has an even number.

Thus, familiarity has been invoked to explain both ‘bad’ and ‘good’ logical reasoning.

---

<sup>5</sup>A tribe in Northern Eurasia.

## Chapter 2

# Syllogisms and beyond

### What is the syllogistic task?

After all the papers that have been published in the psychological literature on syllogistic reasoning, it might be supposed that we know what the task involves—both what it is intended to involve by the experimenters, and what it is interpreted to involve by the subjects. Evidence that this is not the case is revealed by close examination of researchers' own statements as well as of their subjects' responses, as we shall see directly.

In contrast to the many interpretations of the syllogistic *task*, there is at least almost universal agreement about the *logic* of the syllogism.

**Syntax** A syllogism consists of two premises which relate three terms (a, b and c), one of which (the *middle* term, b) occurs in both premises, while the other two (the *end* terms, a and c) each occur in only one premise—a is the end term in the first premise, and c is the end term in the second premise.

There are four *moods* or premise types, distinguished by the quantifiers “all”, “some”, “none” and “some...not”. The quantifiers *all* and *none* are *universal*. The quantifiers *some* and *some...not* are *existential*. There are four possible arrangements of terms in the two premises, known as *figures*, as shown in Table 2.1. We also make use of the term *diagonal figures* to refer to the first pair of figures (ab/bc and ba/cb) and *symmetric figures* to refer to the second pair (ab/cb and ba/bc). Since each premise can be in one of four moods, and each premise pair can have one of four figures, there are  $4 \times 4 \times 4 = 64$  different syllogisms.

Figure Number	diagonal		symmetric	
	1	2	3	4
1st premise	a-b	b-a	a-b	b-a
2nd premise	b-c	c-b	c-b	b-c

Table 2.1: The four figures of the syllogism.

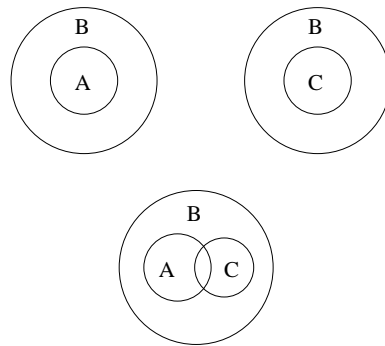
**Semantics** The semantics of the syllogism concerns *types* of individuals defined by combinations of properties. Because there is no identity relation, it is never of logical significance how many of a type of individual exist—hence the talk of *types*. Properties are denoted by the terms of the syllogism—given the term  $a$ , any individual is either an  $a$  or not an  $a$  ( $\neg a$ ). Fully specified types are specified with respect to all three properties, and there are eight such types:  $abc$ ,  $ab\neg c$ ,  $a\neg bc$ ,  $a\neg b\neg c$ ,  $\neg abc$ ,  $\neg ab\neg c$ ,  $\neg a\neg bc$ , and  $b\neg a\neg b\neg c$ .

We can express the semantics of syllogisms in terms of sets of types. *Models* of syllogisms are sets of fully specified types whose existence is consistent with the truth of both premises. Each set can be thought of as a possible syllogistic world described at the level of abstraction with which the syllogism deals. If all the models of a sentence are included in all the models of a pair of premises, then that sentence is a valid conclusion of that syllogism.

Conventionally, in the psychological literature, as in the traditional, but not the modern logical literature, the syllogism is interpreted under the assumption that none of the three sets of things with the properties  $a$ ,  $b$  and  $c$  are empty<sup>1</sup>. Sometimes this assumption is explicitly included in the instructions to subjects, although, as we shall see it is an open question whether they take this instruction on board. This ‘no-empty-sets’ axiom reduces the number of possible models somewhat. A modern logical treatment would make these existential assumptions explicit, separately from its definition of the quantifiers.

**Inferential Structure** Of the 64 syllogisms, under the specified interpretation, 27 have valid conclusions which can be formulated by applying one of the four quantifiers to the two end terms. The remaining 37 syllogisms

<sup>1</sup>There are interesting historical reasons: the distinction between the domain of interpretation of an argument and the universal domain wasn’t clarified until the 20C. If one thinks of properties in the universal domain, it is more plausible to assume that primitive properties are always instantiated somewhere.



All A are B  
All C are B

So, Some A are C?

Figure 2.1: A syllogism and potential conclusion represented by Euler's circles.

do not allow any valid conclusions.

### 2.0.1 Euler's circles

Most diagrammatic methods for solving syllogisms are based on the same spatial analogy—the analogy between an item's membership in a set and the geometrical inclusion of a point within a closed curve in a plane. Euler took this analogy, represented the first premise as a pair of circles, represented the second premise by adding a third circle; and finally read off the conclusion from this construction. So, for example, Fig. 2.1 shows how one might solve the syllogism *All A are B. All C are B.*

This is an interesting example illustrating the inexpressiveness of diagrams. While we may feel that Fig. 2.1 is helpful in clarifying the reasoning, as a method of reasoning it is obviously fallible. Even setting aside the several different diagrams we could have chosen to exemplify the premises, there are several ways of combining the two premise diagrams that we did choose. The conclusion drawn here is true in the combination we chose, but if we had only tried a different combination, we might have found that the conclusion did not hold. Diagrams exhibit specificity. How are generalizations to be extracted from them?

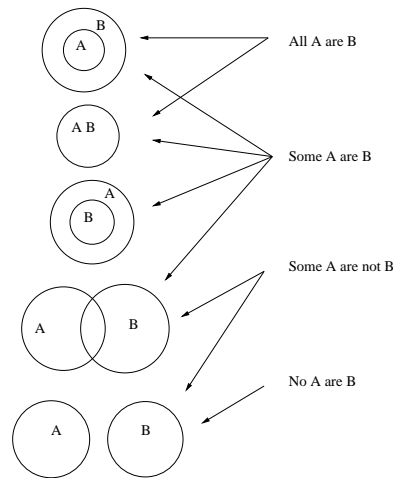


Figure 2.2: The five topological relations between two circles mapped onto the quantified sentences they model—the Gergonne relations.

Figure 2.2 shows the mapping of the sentence forms onto primitive Euler diagrams. The problem of the explosion of combinations of diagrammatic elements is particularly acute in the case of *some* where four of the five topological<sup>2</sup> relations between two circles exemplify the sentence. Imagine having to solve the syllogism *Some A are B. Some B are C* by this method. There are four diagrams of the first premise, four of the second, and several ways of making each combination to derive a composite diagram. There are about 50 diagrams we should consider. In many of these diagrams, it is true that some A are C (or, diagrammatically, that circle A intersects circle C), but not in all. If we chanced to construct a diagram where some A were C, we might make this conclusion without realizing that it did not follow because there were other situations which made the premises both true, but in which this conclusion was nevertheless false.

This problem for diagrammatic methods of reasoning stems exactly from the lack of abstraction inherent in diagrams. This problem has been used by Johnson-Laird to argue for the hopelessness of diagrammatic methods. How could a brilliant mathematician like Euler make such a fundamental

<sup>2</sup>The term ‘topological’ is used here to refer to the possible spatial relations between the circles. It has nothing to do with ‘topology’ in the mathematical sense. In fact, one should not think of the circles as either open or closed sets, because this would lead to a nonclassical semantics for negation.

mistake? How did the poor princess gain anything from this flawed system?

The answers, of course, are to be found in what Euler taught about the *strategy* for choosing which diagrams to use when. Euler taught that one should select what we might think of as the ‘weakest’ case. If we want to represent *Some A are B* we do not choose a diagram where all A are B. If we want to add a third circle to a diagram, we do so in the way that represents the most possibilities—graphically we include the most circle intersections that are consistent with the premises. Although Euler did not burden his princess with adding these strategies explicitly to his notation, it is rather simple to do so. Figure ?? shows four diagrams for representing the four premises.

The crosses mark what are technically known as minimal models. If there is a cross in a minimal subregion of the diagram, then there *must* be something in the world represented with the properties corresponding to that subregion. One thing corresponding to a cross is the absolute minimal ‘furniture’ there must be in a world for the sentence to be true in that world. There may be more, but this much there must be. If there is more than one cross, there is more than one minimal model. It is a property of the syllogism, that minimal models are always single-element models.

The crosses help capture the right strategy for combining diagrams. Figure 2.3 shows the process of solving the syllogism *All A are B. Some C are not B* exploiting the cross notation. The two premise representations are combined by registering the B circles (which are, after all, the same circle). There is then a choice about the relation between the A and C circles. They could be placed in one of three arrangements consistent with the premises. The rule is that the arrangement which creates most subregions is chosen. A simple rule determines whether the crosses persist or are eliminated during this combination process. If a subregion containing a cross is bisected in combining the diagrams, then its cross is eliminated. If not, then the cross persists. Finally, we need to determine what if any conclusion follows from the represented syllogism.

It turns out that every valid conclusion is based on the three-property description of a subregion containing a cross (in the case of Fig. 2.3, the subregion with a cross in the final three-circle diagram corresponds to Cs that are not A and not B). No cross, no valid conclusion. Conversely, if there is a cross remaining in the final diagram, then, with a couple of interesting exceptions which we shall come to directly, there is a valid conclusion. To generate a conclusion, first describe the type of individual corresponding to the cross’s region and eliminate the B term (in the case of Fig. 2.3 this yields ‘things that are C and not A’). Now, if the subregion containing the

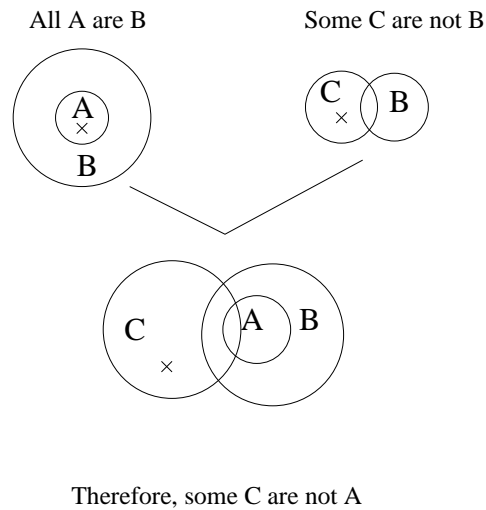


Figure 2.3: An example syllogism solved by Euler's method augmented by the cross notation.

cross is circular, its label becomes the subject of a universal conclusion. Otherwise the conclusion is existential (with *some* or *some\_not*). If the cross is outside one of the two A and C circles, then the conclusion will be negative; otherwise it will be positive. This algorithm will generate all the valid conclusions of the 64 syllogisms. Try some!

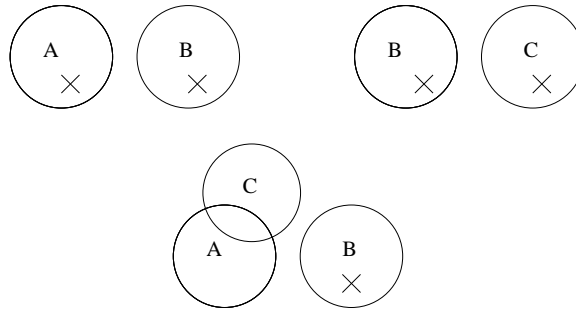
The attentive reader will be asking why this graphical rigamarole leads to the right answer? How is it any more insightful than the medieval logician's mnemonics about Barbara and her exploits? We shall look at the differences between readers who find it satisfying, and those who do not later. But a few hints might be helpful. Combining the circles to give the maximum number of subregions consistent with the premises guarantees the exhaustiveness of the search for kinds of things there may be in a world in which the premises are true. A cross's subregion is guaranteed to have something in it. If the region is bisected during the addition of the third circle, then it is no longer clear which subsubregion the cross should go in, and so neither of the new subsubregions is guaranteed to have anything in it. And so on for the rules about drawing conclusions. If there is no cross left in the diagram there is no type of individual defined for all three properties which must exist. It turns out that the syllogism has the rather special metalogical property that all its valid conclusions follow in respect of such maximally defined individuals. If there is such a cross, then its subsubregion defines the type of thing that

must exist. At least an existential conclusion is therefore justified. The rules about when universal and negative conclusions are warranted are left as homework.

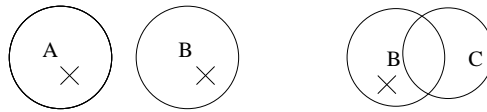
Our cross notation abruptly turns Euler's method from one mired in up to 50 diagrams for a syllogism, to one which provides a one-pass algorithm for solving any problem. The cross notation is an abstraction trick. It allows exactly the abstractions required, but little more. The result is a kind of mechanical calculating device. Imagine the circles as wire hoops of variable size, and the crosses as drawing pins (thumb tacks in American). This mechanical device models the logical constraints of the premises. If a drawing pin prevents the A and C wire hoops from being pulled horizontally apart, then a positive conclusion follows about the fact they must intersect. If the drawing pin prevents the A and C hoops from being pushed into complete correspondence, then a negative conclusion is justified by their non-correspondence. No drawing pin—no conclusion. Euler's genius was to appreciate this correspondence between logical and mechanical constraints.

Euler's method exposes metalogical properties of the syllogism which are important to their implementation in representation systems. As we just noted, conclusions require there to be a cross in the final diagram, and the cross corresponds to a completely described type of individual. The diagrammatic representations make this especially easy to see. The geometrical properties of closed curves (such as circles) mean that every point in a plane is either inside or outside any closed curve. Therefore any cross defines its corresponding type of individual with regard to all three properties. Every valid argument can be made in virtue of a cross in a final diagram. We shall call this property of the syllogism revealed by the diagrammatic system *case identifiability*, there being always a single 'case' that founds a valid argument. Most logical systems do not have this property. A further obvious question is whether the converse of this no-cross—no-conclusion generalization is also true? Does every cross give us a valid conclusion?

For Aristotle's system the answer is no. There are syllogisms which give rise to final diagrams containing crosses which do not allow any Aristotelian conclusion. Figure 2.4 shows two such syllogisms and their diagrams. If we think for a moment about whether there *should* be a conclusion to these problems, we can see that in all cases it is true that *Some not-A is not C*. This somewhat inelegant expression is not included in Aristotle's system, but if we add it, then we can catalogue more valid conclusions. Since the conventional abbreviations for *all*, *some*, *none*, *some\_not* are respectively *a*, *e*, *i*, *o* (from the Latin mnemonics), we shall abbreviate this new quantifier *u*.



No A are B. No B are C. So Some not-A are not C.



No A are B. Some B are not C. So Some not-A are not C.

Figure 2.4: Syllogisms with valid conclusions *Some not-A are not-C*—the ‘u-conclusions’.

It is quite unlikely that Aristotle's omission is to be put down to the inelegance of phrasing of this proposition (in English or in Greek), nor to mere oversight. He had much more systematic reasons for excluding these conclusions which turned out to be at the heart of one of the critical developments of twentieth century logic. Aristotle did not make a distinction between the domain of interpretation of a logic, and the universal domain (everything in the universe—the whole lot). He thought of all sentences being interpreted on the universal domain. This is why the no-empty-properties assumption was more natural for him than for a modern logician. If all sentences are thought of as interpreted on the universal domain, then the proposition that there is something somewhere that does not have properties A and B is fairly vacuous (I say 'fairly'). It amounts only to the claim that not everything does have one or the other property. The development of twentieth-century semantics grew out of the realization that logics *must* be interpreted on local domains, and that the idea of the universal domain is incoherent. Russell's paradoxes, which brought down Frege's anti-semantic approach to predicate logic, were what finally exposed this incoherence at the turn of the twentieth century. This twentieth-century insight into the distinction between local and universal domain is the logical foundation for our cognitive focus on the importance of reasoning externally about systems. The logical discoveries have not yet permeated psychological theories of reasoning.

Quite apart from his metatheoretical reasons, Aristotle would have had some discomfort about admitting these u-conclusion syllogisms as valid because they violated one of his two major generalizations about validity, namely:

1. at least one of the premises must have a universal quantifier for a syllogism to have a valid conclusion
2. at least one of the premises must have a positive quantifier for a syllogism to have a valid conclusion.

Syllogisms with only valid u-conclusions violate this second dictum. These logical observations are not merely historical curiosities. They raise interesting psychological questions about whether logically untrained students have any metalogical, if implicit, grasp of these principles. And does their grasp agree with Aristotle, or with these more recent insights? We return to these questions when some data on how students reason allow logical insights to be put to good psychological effect.

A final metalogical property of the syllogism that is emphasized by Euler's system is what we might call its self-consistency. Euler's system is

self-consistent in that no single diagram can represent an inconsistent set of propositions—the diagrams are models.<sup>3</sup> Strictly speaking, we need to make some assumptions explicit before this statement is true. We have been assuming that Euler’s system represents properties by *continuous* closed curves (ones which are not composed of separated regions); and that distinct closed curves in the same diagram represent different properties. For an example of a discontinuous closed curve, imagine the two circles in Fig. 2.3 interpreted as a single closed curve called A. The possibility of discontinuous curves would be incompatible with the reasoning algorithms. Similarly, in the same figure, if the two circles A and B denoted the same property, then the diagram would express a contradiction. Self-consistency is a consequence of saturated direct interpretation.

## 2.1 Individual differences in reasoning

### 2.1.1 Immediate inferences and the understanding of quantifiers

By an ‘immediate inference task’ we mean a reasoning task in which a subject is presented with a single premise and then has to judge whether a given conclusion follows. In slightly more detail: subjects are asked to assume that the premise is true, and then asked whether the conclusion is to be evaluated as *true*, *false*, or as *can’t tell*. There is no mention of logic in the instructions, and generally the subjects in these experiments have not had any instruction in logic. There is little reason to assume that ‘naive’ subjects, in a decontextualised task such as this, reason according to the canons of logic, and this mainly because in many everyday contexts, logical reasoning is strengthened by pragmatic considerations of the kind highlighted by Grice.

**Information packaging** ‘Some A are B’ is different from ‘Some B are A’ in the sense that these sentences are answers to different questions. 30% of subjects deny that the latter sentence follows from the former. The difference between the sentences is described as ‘information packaging’: the same information is realised (‘packaged’) linguistically in different ways, and this because each packaging serves a different communicative function. If a logically naive subject is asked whether ‘Some B are A’ follows from ‘Some A are B’, he may not realise that he should strip the sentences off their in-

---

<sup>3</sup>Logically, a model of a set of sentences is an interpretation (assignment of content) which makes the sentences all true.

formation packaging, he may not even realise that that is what the graphical representation using Euler circles does.

**Grice and cooperative communication** In *cooperative* communication, the hearer assumes that the speaker is being helpful, and will tell the hearer exactly what he needs to know. By contrast, in *adversarial* communication, such as legal argument, the speaker will tell the hearer what he (the hearer) is forced to admit. The two communicative situations lead to very different notions of ‘follows’, as first pointed out by Grice [37]. Cooperative communication gives rise to *implicatures*, inferences that can be made on the assumption that the speaker is helpful, although not from their logical content. Thus, if the speaker says ‘Some A are B’, the hearer generates the implicature ‘Some A are not B’: for if the speaker actually believed that *all* A are B, he would have said so, being helpful<sup>4</sup>. If the hearer were to generate this implicature in a legal context, he would be laughed out of court, precisely because in this context unstated assumptions are not allowed in inference.

Again, in a decontextualised experiment some subjects may have a cooperative model of communication, and they will exhibit inferences of the type ‘Some A are B’ implies ‘Some A are not B’. It is however not quite clear what counts as an implicature. Consider ‘All A are B’. Should we conclude from this ‘Some B are not A’, or ‘All B are A’? The second inference is more common than the first. The trouble is that there is no natural language quantifier which expresses co-extensiveness of A and B, so that from an utterance of ‘All A are B’ one may conclude either that the converse is false, or that the speaker did not bother to use a clumsy phrase for expressing equality.

**Experimental data** A striking observation is that subjects’ answer-patterns can be classified using relatively few dimensions, roughly corresponding with tendencies toward assuming cooperative communication and information packaging. Let ‘T’, ‘F’, ‘CT’ abbreviate the responses ‘true’, ‘false’ and ‘can’t tell’. Call a subject who tends to answer T or F when the answer should be CT *rash*, and a subject who tends to answer CT when the answer should be T or F *hesitant*. The rash subjects are the ones who tend to apply

---

<sup>4</sup>Note that the hearer also has to assume that the speaker knows what the relations between A and B are! This goes beyond being helpful. The speaker may be helpful and not know that all A are B, without knowing that some A are not B. In this case, the implicature generated is only ‘the speaker does not know that all A are B’.

a cooperative model of communication. Now consider a specific structural feature of an immediate inference, namely whether it conserves the order of subject and predicate (henceforth an *in-place* inference), or whether that order is reversed (an *out-of-place* inference). Both rashness and hesitancy can be further evaluated along the in-place/out-of-place dimension. Thus, a subject who is rash in-place is one who tends to answer T or F instead of the correct CT on immediate inferences which preserve subject/predicate structure. It turns out that no subjects were hesitant on in-place questions, which leaves three properties, and hence eight categories. 94% of subjects fall into one of the following four groups: neither rash nor hesitant on any inference (17%), rash on in-place questions, but neither rash nor hesitant on out-of-place inferences (22%), rash on both kind of inferences, but not hesitant (35%), and hesitant on out-of-place inferences plus rash on in-place inferences (20%). It is also striking how many subjects are rash on in-place inferences: 80%! Thus there is good evidence for the prevalence of a Gricean model of communication. Subjects who are rash on in-place, but not on out-of-place inferences are plausibly 'held in check' by information packaging. Hesitant subjects may be completely spellbound by information packaging; it is difficult to say whether they have any idea of logical validity, since 80% of hesitant subjects are rash on in-place questions.

In conclusion we may say that subjects' inference patterns are determined less by the logical notion of validity than by various aspects of the communicative situation, as hypothesised by the subject.

### 2.1.2 Syllogistic reasoning

When we move from immediate inference tasks to full syllogistic reasoning, the task has changed so much that in principle one cannot expect transfer. That is, one cannot expect that in a syllogism with premises

All A are B  
All B are C

the premises are still taken to imply

All B are A  
All C are B

even when a subject has given evidence of this in the immediate inference task. Thus, we cannot predict offhand that a sizable proportion of subjects will draw the conclusion

All C are A.

One reason that this prediction fails is that the presence of two premises leads to an additional complication, namely the grammatical status of the middle term (B) and the end terms (A and C). In the example above, the end term A is in subject position in the premise but in predicate position in the conclusion, whereas for C it is the other way around. As we have seen there is a tendency to find such re-orderings counterintuitive, so that in fact a subject may be saved from committing a logical error by committing another logical error!

**Factors influencing reasoning** In the above we have identified several factors, such as rashness (both on in-place and out-of-place questions) and hesitancy (on out-of-place questions). The syllogism leads to additional factors of possible importance, such as *grammar*: whether the end terms have different or identical grammatical category, and the distribution of the quantifiers over the premises. For example, one may distinguish the case where an existential quantifier (either *some* or *some...not*) occurs in the first or the second premise, in both or in none; and similarly for the other quantifiers.

One may now try to determine the influence of these factors on various aspects of performance. An interesting aspect is whether a subjects chooses an AC ordering or a CA ordering of the end terms A and C in the conclusion of the syllogism. It turns out that this choice is determined only marginally by validity! A statistical model of the factors affecting conclusion term order shows that more important determinants include:

1. when end terms have different grammatical categories, they tend to preserve those categories in the conclusion
2. each quantifier tends to put its end term in subject position, but quantifiers have greatest influence when in premise 2; for instance *some...not* has no influence in the first premise, but a large influence in the second premise
3. hesitant subjects tend to respond with an AC ordering throughout; by contrast, rashness has no main effect
4. the end term *source-premise* is assigned to the subject position in the conclusion; a source-premise is a premise that provides for the existence of critical individuals

We will now proceed to give an informal description of the statistical model.

**From data description to a model** Stenning & Yule [99] presented a general model of how subjects go about syllogistic reasoning. This model has the virtue that it is rather abstract about the particular representations which subjects use, but it nevertheless provides the best explanation of conclusion term ordering that is available. We first describe that model, and then show how it can be extended to incorporate the individual differences in reasoning as a function of quantifier interpretation described above.

Stenning & Yule's model defines **source** and **non-source** premises of syllogisms, using the concept of a **critical individual**. A critical individual is a type of individual specified with regard to all three properties of a syllogism, which constitutes a minimal model of the problem. So, for example, for the syllogism *All B are A. Some C are B. Therefore some C are A.* the critical individual type is an A that is B and C. The source premise is the premise which establishes the critical individual—in this case the second premise. In the Euler diagram system described above critical individuals correspond to crosses in completed problem diagrams. Problems with valid conclusions always have at least one source premise (sometimes both premises are potential sources).

A general algorithm for determining which premise is validly source is fairly complicated, but one very simple principle gets most cases. For problems with valid conclusions, existential premises are always source. (NB problems with valid conclusions only ever have single existential premises). Once a source premise is identified, a conclusion can be drawn. Again the fully general algorithm is complicated but very simple methods get most problems right.

With this much conceptual framework, the psychological model can be stated succinctly: subjects designate a source premise, attach the end term of the other premise to its end, and delete the middle term. So for the example above, they choose the existential premise as source, attach the end term of the non-source premise to its end (yielding *Some C are B and A*), and delete the B. Stenning & Yule present evidence from a task where subjects have to describe critical individuals and thereby produce a full-ordering of all three terms to get evidence that this model captures a wide range of subjects' conclusion term ordering behaviour. Of course, the model abstracts over issues about how source premises are chosen. This model predicts that the middle term should often come first in a description of

the critical individual for a problem. This is the pattern Stenning & Yule observed—in fact the commonest term order of all is BAC. Mental models theory predicts that middle terms always should come between the end terms though it never sought any evidence for this claim it explained the result on the basis of certain memory theories. In fact mental models' theory's prediction appears to be false.

So the Stenning & Yule model conceives of syllogistic reasoning as subjects choosing a source premise which founds a description of a critical individual. It is this model which we now expand in order to capture the individual differences observed with regard to how interpretation determines reasoning.

**Incorporating interpretation factors into the reasoning model** In the statistical model of term-order, the individual differences mostly revolve around the negative quantifiers Some-not and No, particularly the latter. We need to understand the relations between these observations embodied in a statistical model of the data, and the Stenning & Yule process model. A key to understanding this relation is that No-premises are never source.

In order to get a better handle on the complex data, we represent them as pairs of problems which are related by premise-reordering (e.g. *All A are B. All B are C.* and *All B are A. All C are B.* are such a pair. The 27 problems with valid conclusions are composed of 13 such pairs and one singleton which is symmetrical about this ordering. Now, the Stenning & Yule model of source-identification provides a criterion for defining a canonical and non-canonical ordering of each pair. If we order the problem pairs so that problems with existential before universal, and then positive before negative problems, then the source premise will be first and the non-source second. We call the first problem 'canonical' and the second 'non-canonical'. There is one All/All problem that has to be ordered by its grammar (*All AB. All BC* precedes its counterpart. This definition of canonical problems according to the Stenning & Yule model means No-premises are always second premises.

We can ask for any pair of problems and any group of subjects whether they find the canonical problem or its non-canonical counterpart easier. If there is an effect of canonicity, then it means that premise-order is playing a role in subjects' strategies for finding source premises. From the statistical model of conclusion term ordering, we know that hesitant subjects are more influenced by term-order. They are duly more affected by canonicity, showing an advantage in reasoning accuracy for canonical problems. There

are some interesting wrinkles—particular problems where the simple source-finding algorithm tries to put negated predicates into subject position, and the canonicity effects are reversed for these problems for hesitant subjects, whereas rash-out-of-place subjects find both problems hard. Hesitant subjects appear to using premise-ordering to resolve the problem and this gets the correct answer on the non-canonical problem.

In summary, the interactions in the term-order model between quantifier with interpretation variables show that hesitant subjects lean on premise-order (amongst other factors) to determine their choice of source premise, whereas rash-out-of-place subjects are independent of premise order in their choice of source premise. This is as we would expect if hesitant subjects are more susceptible to the information packaging of problems. Analysis of subjects' accuracy profiles across the range of valid problems shows that these factors influencing term-order also affect accuracy of reasoning in exactly the ways one should expect. These factors affecting term-ordering are plausibly instrumental in determining reasoning.

One obvious speculation emerging from this work is that hesitant subjects are using 'verbal' representations and rash-out-of-place subjects are using 'visual' representations (such as Euler's circles which strip out information packaging). Ford [28] has studied subjects whose reasoning protocols indicate 'verbal' or 'visual' strategies for syllogism use. We can reanalyse her data and show that, as one would predict, her verbal subjects find canonical problems easier than non-canonical problems, and her visual subjects the reverse. However, when we compare our hesitant and rash-out-of-place subjects' profiles of problem accuracy with her subjects' profiles, the patterns are obscure. Her subject groups show strong contrasts with both our groups. So there is some evidence that canonicity plays the predicted roles in visual and verbal thinking, but there is more to our interpretation differences than visual and verbal representations (which is not surprising). It is perfectly possible to vary cooperativeness of model with type of representation. It will take further work to resolve the relation between the interpretation variables and the modality of representation involved.

To tie this back to the relation between interpretation and reasoning, we have shown that there are highly systematic patterns of naive interpretation of quantifiers which can be understood as affected by the choice of cooperative or adversarial models of communication in the task, and by differences in susceptibility to information packaging of propositions. These systematic patterns of interpretation can be shown to be instrumental in determining subsequent patterns of reasoning provided they are interpreted in terms of a theory of the relation between models of communication and of information

packaging.

These results raise some important questions for pragmatic theories such as Grice’s, and for the relation between information packaging and language use more generally. The topic of information packaging does not arise in Grice’s theorising. He does not consider whether a statement such as ‘Some A are B’ generates the implicature that ‘Some B are not A’ (the ‘out-of-place’ implicature). If we ask why not, it is obvious that in typical decontextualised natural language examples, the propositions are already packaged into referring expressions and properties. Learning elementary logic is learning to stand back from this pre-packaging and think in terms of local interpretations in which both quantified expressions and predicates designate sets. When this is done, both in-place and out-of-place implicatures become relevant. Standard linguistic methodology only studies how information is distributed across grammatical structures *after* decisions about subject and predicate, term and predicate, etc. have been imposed.

## 2.2 ‘Mental models’

This approach can be seen as a development of the founding idea of Kamp’s *Discourse Representation Theory* (cf. [62] and [63]): a piece of discourse is interpreted in a mental representation (‘model’), which is itself embedded in ‘the world’. Conversely, mental models can be generated by perception, and can be used as conceptual structure underlying language.

**Experimental evidence** It is possible to investigate experimentally whether a particular piece of discourse is stored in memory *verbatim* or in the form of a situation described by the discourse.

1. Bransford, Barclay and Franks [11] provide evidence that recalled sentences are inferences from explicitly presented material—this can easily be explained if people construct a model of discourse, and ‘read off’ what is true there. E.g. when given the sentence

Three turtles rested on a floating log and a fish swam beneath them

subjects later confused it with

Three turtles rested on a floating log and a fish swam beneath *it*.

[But can't the inference be explained by an application of general knowledge to verbal material?]

2. Recall of a piece of discourse is facilitated if the discourse determines a unique model. Cf.

The spoon is to the left of the knife.  
 The plate is to the right of the knife.  
 The fork is in front of the spoon.  
 The cup is in front of the knife.

and

The spoon is to the left of the knife.  
 The plate is to the right of the spoon.  
 The fork is in front of the spoon.  
 The cup is in front of the knife.

Stenning [92], [93]; Mani and Johnson-Laird [70] found in several experiments that (a) the verbal recall of indeterminate description is *better* than that for determinate description, (b) the 'gist' of a determinate description can be recalled much more accurately than that of an indeterminate description. These results lend support conclusion that both models (for determinate descriptions) and verbal material (for indeterminate descriptions) are stored in memory.

**What is a mental model?** That is, does there exist a principled way to distinguish between mental models and linguistic representations? Stenning [94] makes a distinction between *direct* and *indirect* representations, the former do not have syntax. The distinction can be illustrated using the concept of a first order model. Such a model can be completely specified by indicating which tuples of elements belong to which predicates. Equivalently, it can be specified by its *atomic diagram* when all elements of the model are named by constants. The atomic diagram is a set of atomic formulas without further structure: this is typical for a direct representation, where no further parsing is necessary. Alternatively, a (finite) model could be completely specified by a first order sentence; now all the work is in the parsing. Intermediate cases can occur; for instance in Feferman's type-free logic with truth predicate, one may specify the intended model via the truth predicate, and apply the truth axioms to get at the model itself. The 'mental models' in Johnson-Laird's sense appear to be such an intermediate case: partly specified at the atomic level, partly specified via syntax.

**Mental models and reasoning** An important claim of mental models theory is that reasoners are sensitive to the *content* more than the *form* of premises when they make inferences. Here is an example offered in support of this claim; examples of this kind will be discussed at greater length in the section on the 'suppression effect'. Compare the following two conditionals.

- (1) If patients have cystitis, they are given penicillin.
- (2) If patients have cystitis and are allergic to penicillin, they are given penicillin.

People may assent to (1), but refuse to assent to (2), even though (2) follows from (1) in classical logic. This is taken to be an argument against 'mental logic', which would be wedded to applying a rule blindly, i.e. by looking only at the *form* of the statements involved. A criticism of this view of the mental models theorists is that logical form is not something that can be read off from a sentence, is not an intrinsic property of a sentence, but is imposed upon it by the person trying to understand the sentence. The logical form chosen may depend on the whole discourse, not only on the sentence itself. This point will be elaborated below, e.g. in section 3.2.

**The 'algorithm' of the mental models theory** The theory tries to steer a mid-course between rule theories, which purportedly fail because of content effects such as the above, and 'no logic' theories, which go counter to the observation that subjects are able to make valid deductions when pressed. Johnson-Laird claims that the process of deduction goes through the following stages:

1. reasoners use general knowledge to interpret the premises given and as a result come up with an internal model
2. the model is scanned for interesting information not yet explicitly contained in the premises; if there is such information, this is offered as a putative conclusion, if not, the conclusion is that nothing follows
3. reasoners now try to validate the putative conclusion by looking for alternative models of the premises where that conclusion is false.

"If there is such a model, the prudent reasoners will return to the second stage to try to discover whether there is any conclusion true in all the models they have so far constructed. If so, then it is necessary to search for counterexamples to it, and so on, until the set of possible models

has been exhausted. Because the number of possible mental models is finite for deductions that depend on quantifiers and connectives, the search can in principle be exhaustive. (Johnson-Laird & Byrne [57, 35-6])

The reference to finiteness here is incomprehensible (unless it refers to syllogistic reasoning, for which small models suffice), but it highlights one of the aims of mental model theory: to construct a ‘psychologically viable’ alternative to logic, by eliminating the latter’s infinities. Johnson-Laird likes to quote a remark due to Barbara Partee to the effect that ‘an infinite set is far too big to fit inside anyone’s head’. This depends, of course: infinite sets may have compact finite descriptions. In any case

The psychological theory therefore assumes that people construct a minimum of models: they try to work with just a single representative sample from the set of possible models, until they are forced to consider alternatives. (Johnson-Laird & Byrne [57, 36])

It remains to come up with an algorithm that chooses the ‘single representative sample’ and that guides the search for alternatives. Presentations of this subject are notoriously hazy (see e.g. Wilfrid Hodges’ criticism of ‘mental models’ in the issue of *Behavioural and Brain Sciences* 1993 devoted to a discussion of [57], but we will try to be charitable.

In this section we consider the mental models approach to syllogisms only; below we extend these considerations to full predicate logic. Take a premise like ‘all of the athletes are bakers’. The ‘initial model’ would be a partial structure containing elements  $a$ ,  $b$  (for athletes and bakers, respectively), together with some indication of which predicates can be further extended, and which cannot. Johnson-Laird and Byrne use the notation

$$a \quad b$$

to mean that  $a$  is also a  $b$ , and

$$[a] \quad b$$

$$[a] \quad b$$

$$\dots$$

for the situation in which the predicate ‘athlete’ cannot be further extended (beyond the two athletes explicitly listed), in contrast to ‘baker’ or indeed

any other predicate. Such a partial structure (called *implicit* model by Johnson-Laird and Byrne) can be 'fleshed out' to obtain a fully classical structure (called *explicit* model). For example, the above partial structure could be fleshed out to

$$\begin{array}{ll} [a] & [b] \\ [a] & [b] \\ [\neg a] & [b] \\ [\neg a] & [\neg b] \end{array}$$

By  $\neg a$  an element is meant that is not an athlete.

A few remarks are in order.

1. The number of elements in a model is arbitrary, but in (syllogistic) practice a small finite number. (In fact, a simple tableau-argument shows that one never needs more than three elements for arguments of syllogistic form. The maximum number of three is reached for the invalid syllogism having two existential premises and a universal conclusion.)
2. Models in this sense contain some logical devices, such as  $\neg$  and the square brackets, which more or less represent the universal quantifier as restricted to finite sets (but see below). They are thus unlike a logician's models in that some amount of syntax is incorporated. We will explain later why this is significant.

On the basis of the partial model, a conclusion is formulated, which (ideally) is then tested for validity by searching for counterexamples. This can be done either by fleshing out the initial partial model, or by moving to a different initial model. For example, although the initial model validates 'all the bakers are athletes', the fleshed out model refutes this. The claim is that the assumption of this procedure explains all aspects of reasoning performance, including common errors and response times.

We proceed to give a few initial, implicit models for other syllogistic premises (taken from Johnson-Laird and Byrne [57]).

Some of the athletes are bakers.

$$\begin{array}{ll} a & b \\ a & b \\ \dots & \end{array}$$

None of the athletes is a baker.

[*a*]

[*a*]

[*b*]

[*b*]

Here, the empty space to the right of the *a* indicates, that there is no *b* such that *a* is *b*; and likewise for the empty space to the left of the *b*.

Some of the athletes are not bakers.

*a*

*a*

*a* [*b*]

[*b*]

...

The principles guiding the construction of these minimal models are somewhat unclear. Note that the implicit model for the first sentence is compatible with ‘all athletes are bakers’, whereas the implicit model for the third sentence is not compatible with ‘none of the athletes are bakers’. No reason for this asymmetry is given. It is also unclear why some of the lines are reduplicated.

We so far considered arguments with a single premise only. When two premises have to be combined, complications arise. Consider

All the athletes are bakers.

All the bakers are canoeists.

A partial model for the first premise is

[*a*] *b*

[*a*] *b*

...

and for the second premise

$$\begin{array}{l} [b] \quad c \\ [b] \quad c \\ \dots \end{array}$$

How are these to be combined? The intended meaning of the square brackets excludes the fused partial structure

$$\begin{array}{l} [a] \quad [b] \quad c \\ [a] \quad [b] \quad c \\ \dots \end{array}$$

because then suddenly 'all bakers are athletes' would be true in all fleshed out models. Johnson-Laird and Byrne opt for the representation

$$\begin{array}{l} [[a] \quad b] \quad c \\ [[a] \quad b] \quad c \\ \dots \end{array}$$

which is supposed to mean that all *pairs*  $[[a] \quad b]$  have been listed explicitly. Notice however that we have now moved to pure syntax. Whereas '[a]' could be interpreted as a meta-language device to indicate that all elements  $a$  are explicitly listed, the iteration of the square brackets requires their presence in the object-language, and even then their interpretation is not very perspicuous. Johnson-Laird and Byrne offer the interpretation ' $a$  is exhausted with respect to  $b$ , and  $b$  is exhausted with respect to  $c$ ', but this gloss destroys the original meaning of  $[a]$  as being exhausted *tout court*. In logical terms, the difficulty can be put thus. If we read the collection of  $[a] \quad b$  as an abbreviation for  $\text{forall}x(Ax \rightarrow Bx)$ , then the collection of  $[a] \quad [b] \quad c$  would probably mean something like  $\forall y(\text{forall}x(Ax \rightarrow Bx) \rightarrow Cy)$ , which is equivalent to  $\text{forall}x(Ax \rightarrow Bx) \rightarrow \forall yCy$ , and this is not what we want. It almost goes without saying that formation rules for the '[...]'-construct are not given.

It is perhaps best to think of the square bracket notation in graphical terms. Consider the expression  $[[a] \quad b] \quad c$ . This is analogous to an Euler circle representation of sets  $A, B, C$ , with  $A \subseteq B \subseteq C$  and distinguished elements  $a \in A$ ,  $b \in B$ ,  $c \in C$ . The added twist is that extensions of the predicates must leave the arrangement of the Euler circles intact. Interestingly, however, it is much less natural to think of Euler circles as representing a partial

model, since the circles neatly abstract from the number of elements in the model.

The next question is how to read off a conclusion concerning athletes and canoeists from the combined model. This is done by stripping the model of its square brackets; we then see that all athletes are canoeists, and no extension of the partial model can invalidate this conclusion. Let's see how this works in a more difficult case.

All of the bakers are athletes.

None of the bakers is a canoeist.

A partial model for the first premise is

[*b*] *a*

[*b*] *a*

...

and for the second premise we get

[*b*]

[*b*]

[*c*]

[*c*]

The combined model is something like

[*a* [*b*][

[*a* [*b*]]

[*c*]

[*c*]

...

which is meant to suggest that triples like

*a b c*

are excluded. This model supports the conclusion

None of the canoeists is an athlete.

which is apparently not uncommon. A fleshed out model such as

[a [b]]  
 [a [b]]  
 a [c]  
 a [c]  
 ...

invalidates this conclusion, and supports

Some of the athletes are not canoeists.

It will be clear from the above exposition that Johnson-Laird and Byrne's claim to have provided an 'algorithm' for deduction should be taken with a grain of salt. The choice of the initial implicit model is not completely determined, neither is the fleshing-out. However, with a number of more or less arbitrary assumptions added, the above can be cast in the form of a computer program, for which see [www.cogsci.princeton.edu/~phil](http://www.cogsci.princeton.edu/~phil), or the recent paper by Bara et al. in *Cognitive Science*.

**Predictions** It is the aim of mental model theory to be able to predict both correct performance when it occurs, and frequently occurring errors. One important determinant of successful performance is taken to be the number of models that must be constructed before a valid conclusion is obtained. Thus, for example, the syllogism with premises

All the athletes are bakers.  
 All the bakers are canoeists.

is supposed to require only one model, in the sense that 'fleshing out' does not change the conclusion read off from the implicit model. This is supposed to be in contrast to the situation for the syllogism

All of the bakers are athletes.  
 None of the bakers is a canoeist.

As indicated above, Johnson-Laird and Byrne claim that the conclusion read off from the implicit model is

None of the canoeists is an athlete.

However, the valid conclusion

Some of the athletes are not canoeists.

can also be read off from the implicit model. It will also not do to say that people read off the strongest possible conclusion from an implicit model, because in that case the premises

All the athletes are bakers.  
All the bakers are canoeists.

would lead to the initial conclusion

Athletes are bakers and conversely.

This conclusion would of course be invalidated by the fleshing out of the implicit model, but then it is difficult to see what is meant by saying that this syllogism needs only one model.

### 2.2.1 Full predicate logic

Syllogistic reasoning can be facilitated by graphical representations—which is not to say that all reasoners proceed in this way. For full predicate logic this option is no longer open. That is, there exist graphical representations of predicate logic in terms of so called Peirce diagrams<sup>5</sup>, but these require a lot of logical pre- and postprocessing. Binding and instantiation of variables now come to the fore. How can we do this?

**Exercise** Johnson-Laird and Byrne consider the following result to be striking evidence against rule theories of reasoning. They first presented adults with the premises

None of the painters is in the same place as any of the musicians.  
All of the musicians are in the same place as all of the authors.

They then asked the subjects to draw a conclusion from these premises. The majority of the subjects drew the conclusion

None of the painters is in the same place as any of the authors.

However, when the premises presented were

---

<sup>5</sup>Although a case can be made for construing Peirce diagrams as diagrams of *sentences* rather than of situations.

None of the painters is in the same place as any of the musicians.  
 All of the musicians are in the same place as some of the authors.

only a few subjects drew the valid conclusion

None of the painters is in the same place as any of the authors.

Assume that the predicates mentioned are nonempty. Suppose the rule theory adopted is some form of natural deduction. Can you find a plausible difference in complexity between the two derivations? Setting deduction aside, is there a difference in linguistic complexity between the two sets of statements?

**Mental models theory for multiple quantifiers** Consider again the premises

None of the painters is in the same place as any of the musicians.  
 All of the musicians are in the same place as all of the authors.

Johnson-Laird and Byrne offer the following as an initial model of the first premise:

$$| [p] [p] [p] | [m] [m] [m] |$$

where the vertical lines separate the different places. Note that many more structurally different models would be possible, allowing more locations than just two. The addition of the second premise eliminates those possibilities and gives

$$| [p] [p] [p] | [m] [m] [m] [a] [a] [a] |$$

This model validates the conclusion

None of the painters is in the same place as any of the authors.

In this case there is no structurally different alternative model of the premises, hence *a fortiori* not one which negates the conclusion.

In the case of the premises

None of the painters is in the same place as any of the musicians.  
 All of the musicians are in the same place as some of the authors.

the mental models analysis goes as follows. The first premise is again represented as

$$| [p] [p] [p] | [m] [m] [m] |$$

, but according to Johnson-Laird and Byrne the effect of adding the second premise is the model

$$| [p] [p] [p] | [m] [m] [m] a a |$$

i.e. a model where the  $a$ 's are not exhaustively represented. This model again validates the conclusion

None of the painters is in the same place as any of the authors.

but in this case there is an alternative model having only two locations which invalidates this conclusion, namely

$$| [p] [p] [p] a | [m] [m] [m] a a |$$

This model supports the conclusion

Some authors are in the same place as all of the painters.

which is apparently drawn by some subjects, even though it is invalidated by the first model. A valid conclusion is

Some authors are not in the same place as any of the painters.

but according to Johnson-Laird and Byrne subjects have difficulty in coming up with this conclusion because the premises allow structurally different models.

How to make sense of this claim? Recall that in mental models theory logical reasoning is conceived of as a two-tiered process: reading off a conclusion from a 'minimally informative' model of the premises ('context of discovery'), followed by a search for counterexamples to validate the conclusion ('context of justification'). The latter step is definitely more difficult in the case of the second argument, because the possible models are structurally dissimilar.

Even assuming that subjects understand what is meant by the instruction to 'draw a logically valid conclusion', difficulties with a particular syllogistic form can occur for two reasons in a mental models approach: (1) a subject picks a wrong initial model, or (2) reads off an invalid conclusion, and does not exhaustively search for counterexamples. If subjects would have a systematic procedure to look for counterexamples such as semantic tableaux [...]

**Conclusion: the logic of mental models** The confusions inherent in mental model theory are nicely illustrated (albeit unwittingly) in the following quotation from Ruth Byrne [12, p. 78]:

The model-based theory assumes that once one has an adequate explanation of comprehension, there is no need for the mobilisation of any elaborate machinery for reasoning, neither of syntactic rules proposed by formal theories, nor of domain-specific rules favoured by pragmatic theories. On the contrary, reasoning depends on a search for counterexamples to conclusions, but ordinary individuals do not have a simple deterministic procedure for making such a search (cf. Newell and Simon [74]).

The claim is thus that mental models theory is 'semantic' in the sense that it rests completely on a theory of natural language understanding; this would render the application of rules superfluous. But then, what is involved in natural language comprehension? The assignment of an implicit mental model obviously does not depend on the full semantic content of a sentence such as 'all athletes are bakers', but it depends on the *logical form* of this sentence. This is what allows Johnson-Laird to represent the sentence 'all athletes are bakers' in the form

$$\begin{array}{l} [a] \quad b \\ [a] \quad b \\ \dots \end{array}$$

abstracting from all other information.

In fact, an important component of comprehension is precisely assigning logical form. Once a logical form has been chosen, representing this as a mental model in the sense of Johnson-Laird is a triviality, because, due to the incorporation of logical devices such as negation and the universal quantifier, such a mental model is a notational variant of a sentence of predicate logic. Thus, there is nothing inherently semantic in this notion of a mental model. In this respect it should be sharply distinguished from ordinary first order models, which can be characterised by sets of *atomic* sentences.

A reasoner now supposedly 'reads off' a purported conclusion from a mental model constructed from the premises. Since the mental model is essentially a sentence in a formal language, this involves checking whether

one sentence follows from another, and this is definitely not ‘reading off’ when the sentences involved are complex. Similarly, checking whether the conclusion is true in other models of the premises, again involves determining whether one sentence follows from another. Thus, reasoning processes are involved which are not described by the mental models theory.

All this can be illustrated in a more perspicuous manner with the help of semantic tableaux, which avoid the various red herrings of mental models theory. This proof procedure is semantic in the sense that it provides a systematic way to construct a first order model which verifies the premises and falsifies the conclusion of an argument. Models constructed in this manner are true models in that they are fully specified by a domain and relations (and/or functions) on that domain; the model does not incorporate syntax. But the construction itself proceeds stepwise and is fully guided by logical rules of the type ‘if  $\forall x\varphi(x)$  is false, then there must be some new individual  $a$  such that  $\varphi(a)$  is false’. These rules are very similar to the rules proposed by the rule theorists; indeed, there exists a canonical way to transform a tableau-proof into a natural deduction proof. A mental models theorist might object that even so, there still is a distinction between rules operating on formulas, and rules guiding the construction of representations or models. This distinction is spurious however, because semantic tableaux can be viewed equivalently as fully syntactic procedures, by treating the elements introduced in the course of the construction simply as constants of the language.

It will be clear from the above exposition that Johnson-Laird and Byrne have inadvertently tried to reinvent logic, while claiming to provide an alternative. In itself that is already an interesting result. Stripped of its extravagant claims and notational muddle, the mental models theory embodies a simple logical point. What the mental model theory has in common with Kamp’s ‘Discourse representation theory’ is the emphasis on partial models. Now partial models do not seem to be very helpful in the case of syllogisms, because the full models are so small in any case, but they may be helpful for more advanced arguments.

## Chapter 3

# Propositional logic—the easy cases

There have been three main experimental paradigms in studying propositional reasoning. The first paradigm consists in showing a subject a set of statements to be interpreted as premises, and asking her to evaluate whether another statement is or is not a valid conclusion from these premises. Variants of this paradigm include asking the subject to choose from a set of possible conclusions, or to come up with some conclusion. Tasks of this type are generically known as inference tasks.

The second paradigm tests subjects' understanding of propositional connectives, by asking them either to construct a truth table, or to evaluate a given (line in a) truth table. These type of tasks, known as the construction task and evaluation task respectively, are of course (hopelessly) wedded to classical logic.

The third paradigm combines logical reasoning with reasoning toward a decision. The most famous of the tasks is the *Wason selection task*, also known as the *4 card task*.

### 3.1 Inference tasks

We will only discuss inference concerning conditionals, to give a flavour of the results obtained in this area, but there have been many interesting results on the other connectives as well.

The simplest kind of conditional inferences have one conditional premise, and one premise which is an atom or the negation of an atom. The four possible argument patterns are traditionally known as

1. *modus ponens* (MP):  $p, p \rightarrow q / q$
2. *modus tollens* (MT):  $\neg q, p \rightarrow q / \neg p$
3. *denial of the antecedent* (DA):  $\neg p, p \rightarrow q / \neg q$
4. *affirming the consequent* (AC):  $q, p \rightarrow q / p$

The first two are classically valid, the latter two classically invalid.

There exist many studies which investigate the frequencies with which people draw these inferences, and the results are sometimes surprising. Typically, MP is judged to be valid by 90% or more of subjects. For MT this can drop as low as 41%, although a rate as high as 81% has been reported. The rate for the endorsement of the (classically invalid) DA varies between 20% and 75%, as does the rate for endorsement of AC. E.g. in one experiment subjects scored 98% for MP, 81% for MT, 48% for DA and 54% for AC. Another experiment showed a somewhat different pattern: 95% for MP, 62% for MT, 51% for DA and 36% for AC.

It is rather hard to interpret these data. What is clear is that MT is less easily available than MP; indeed, a typical answer is that ‘nothing follows’, instead of ‘ $\neg p$ ’. One reason why interpreting the data is so hard, is that the results obtained change when the  $p, q$  in the conditional premise are replaced by negated atoms. Here, it is important to be precise about what negation means. One may use negation only in explicit form, as in the following argument pattern

If the letter is not G, then the number is 9.  
The number is not 9.  
Hence, the letter is G.

The other possibility is that negation is also used in antonymic form, as in

If the letter is not G, then the number is 9.  
The number is 6.  
Hence, the letter is G.

Apparently, most studies have been done with the former type of pattern, although it is known that reasoning performance changes depending on the type of negation used. In any case the results obtained are very striking. For instance, performance on MT in conditionals with negated antecedents ( $\neg p \rightarrow q, \neg p \rightarrow \neg q$ ) drops, in one case to as low as 12%. It is easy to make sense of this logically: why would subjects accept double negation elimination?

Another observation is that the rate of DA answers drop appreciably (even 12% has been recorded) if the conditionals have negated conclusions. Note that this again need not wonder a logician, because double negation elimination is involved. It is harder to explain why AC drops (e.g. to 31%) when the conditional is of the form  $p \rightarrow \neg q$ . An explanation that has often been voiced is that people tend to interpret the conditional as a biconditional, which would then go some way toward explaining the high rate of AC and DA answers in the task involving plain  $p \rightarrow q$ . But even if that is so (which we doubt), it remains to be explained why the biconditional reading is less available here.

In the above we have summarised some of the results in the literature. We have done so without looking at the details of the experiments, although as we shall see below when discussing the selection task, task-specific instructions can have a large influence on the answer chosen. We do want to make one comment on the literature though. Pollard and Evans found in the results on conditionals with negated constituents evidence of a ‘negative conclusion bias’, that is, a tendency for subjects to sanction inferences with negative conclusions, regardless of the logical situation. ‘Bias’ is a supposedly neutral term, which denotes a systematic, but logically irrelevant feature of the data. The negative conclusion bias would be explained by a ‘caution heuristic’. Typically, nouns and adjectives single out a small part of the world. This means that a statement in subject–predicate form such as ‘A is B’ is easy to falsify, but not easy to verify. Precisely the opposite is the case for a statement ‘A is not B’, so a conclusion of this form is more ‘cautious’. Clearly, if subjects would employ a caution heuristic, they would not engage in logical reasoning from premises.

By contrast, a logician would point out that there is a reason *within logic itself* why in several cases it is difficult to produce a positive conclusion: the problem with double negations. This explanation does not cover AC for the conditional  $p \rightarrow \neg q$ , but it does cover the other cases. We shall see when discussing the selection task that there is strong evidence that subjects do indeed have difficulty with the inference from ‘not–false’ to ‘true’. There is thus a contrast between the two explanations: the ‘caution heuristic’ explanation apparently covers all the data, but is shallow in the sense that it has little to do with the task at hand. The logical explanation leaves the AC case unexplained, but is much deeper in that it connects performance of the task to the semantics of natural language. That being said, the two explanations need not be mutually exclusive: the first explanation could be true of subjects who do not really engage with the task (this is all too common!), whereas the second explanation might be relevant when subjects

do engage with the task.

### 3.1.1 Children's propositional reasoning

MP is observed quite early in children. As usual the percentages differ from experiment to experiment, but a rate of 74% for 6 year olds has been reported. Interestingly, the rate of MT seems to increase to around 74% for 10–12 year olds, only to decrease again for older children. The rates for endorsement of DA and AC can be very high; e.g. for 13 year olds, rates of 95% (DA) and 88% (AC) have been reported.

These data raise the question whether children actually understand the conditional nature of *if*. A very interesting experiment to test this is one due to O'Brien et al. Children are presented with four toy boxes containing a (toy) cat and a banana, a (toy) dog and an orange, a (toy) dog and an apple, and a (toy) horse and an apple. The children have to evaluate the conditional 'if there is an apple in the box, then there is a horse'. This conditional is of course false in the set up described, but only 5% of the 7 year olds and 15% of the 10 year olds judged this sentence to be false<sup>1</sup>. An explanation commonly offered is that children read this conditional as 'there is a box which contains an apple and a horse', which is true here. It has also been claimed that this interpretation would explain the results concerning the four conditional inference types, but it is hard to see how this would go. The interpretation does get some support from the observation to be discussed later in connection with the Wason selection task, that adults may also exhibit a conjunctive interpretation of the conditional.

If children would indeed have a conjunctive reading of the conditional, this would mean that their conditional is not hypothetical, whereas for 'us', of course, hypotheticality is among the prime functions of the conditional. This question has been studied from another angle, language acquisition, by Bowerman [8]. It is a fact that conditionals appear fairly late in children's speech, much later than other conjoining items such as *and*, *when*, *because*, *so* . . . . In principle, three kinds of factors could be responsible for this delay: cognitive complexity, grammatical complexity, and pragmatic need, or absence thereof. E.g. the notion of a state of affairs obtaining hypothetically could be too complex to grasp for very young children. Interestingly, however, markers of hypotheticality appear long before conditionals. A pragmatic factor could be that before a certain age there is simply no need for children to employ *if then*, because there are other constructions that can

---

<sup>1</sup>As opposed to a still strikingly low 75% for adults.

do duty for the conditional. For example, a threat such as ‘if you break this I’ll hit you’ can be formulated equivalently as ‘don’t break that or I’ll hit you’. However, this seems to go against the observation that such threats are not among the earliest conditionals produced.

### 3.2 ‘Suppression effect’ and nonmonotonicity

The following experiments due to Ruth Byrne were originally designed to distinguish between mental logic and mental models. Her argument was as follows. People apparently make systematic errors in reasoning (e.g. St Augustine commits ‘denial of the antecedent’<sup>2</sup> in his Commentary on Psalm 54: ‘Si enim justus est qui ex fide vivit, iniquus est qui non habet fidem’). A theory of reasoning should explain also why people make mistakes, and the question then arises whether there are also formal rules corresponding to fallacies. Rule theorists generally deny this and claim that fallacies can often be viewed as Gricean implicatures or invited inferences. For example, St. Augustine would most likely *equate* being just with having faith, in which case his inference follows. A touchstone for such an interpretation would be the possibility to suppress the invited inference when it is clear that the conditional cannot be read biconditionally. For instance, in a situation where a signal may be activated by two switches, people do not make the AC inference

If switch 1 is turned on, the signal is on.  
The signal is on.  
Switch 1 is turned on.

In this case the AC inference is blocked, and this seems to show that there is no formal rule corresponding to AC; for, as the argument goes, if there is a rule it will be applied no matter what. However, Byrne now attempts to turn the tables on the rule theorists by showing that valid inferences can also be suppressed!

The first experimental condition replicates a standard design. Consider the premises

If she has an essay to write she will study late in the library.  
She has an essay to write.

---

<sup>2</sup>Henceforth we shall use the following abbreviations for argument patterns: MP for *modus ponens*, MT for *modus tollens*, DA for *denial of the antecedent* and AC for *affirmation of the consequent*.

Does it follow that she will stay late in the library? Subjects overwhelmingly say ‘yes’, which might seem comforting, were it not for the by now familiar fact that in case of the following argument

If she has an essay to write she will study late in the library.  
She will study late in the library.

71% of Byrne’s population draw the conclusion ‘She has an essay to write’. Likewise, if the premises are modified to

If she has an essay to write she will study late in the library.  
She doesn’t have an essay to write.

then 46% draws the conclusion ‘She will not study late in the library’.

Now suppose that (following Romain et al. [83]) a premise is added, for example as follows

If she has an essay to write she will study late in the library.  
If the library stays open then she will study late in the library.  
She has an essay to write.

Interestingly, the conclusion ‘She will study late in the library’ is drawn by only 38% of the subjects<sup>3</sup>, and the two ‘fallacies’ are committed by 54% and 63% of the subjects, respectively. For the latter, the corresponding argument patterns are

If she has an essay to write she will study late in the library.  
If the library stays open then she will study late in the library.  
She will study late in the library.

and

If she has an essay to write she will study late in the library.  
If the library stays open then she will study late in the library.  
She doesn’t have an essay to write.

Even more interesting is the observation that this pattern of results depends upon the kind of premise added. If the premises are given as

---

<sup>3</sup>There were 8 subjects in all, each doing three arguments of this type. This design is not beyond criticism, since the arguments cannot be considered to be independent. Dieussaert et al. [22] replicate this experiment with much more statistical power and obtain the same pattern of results as Byrne did.

If she has an essay to write she will study late in the library.  
 If she has some textbooks to read, she will study late in the library.  
 She has an essay to write.

then modus ponens is restored, but the fallacies are almost eliminated. That is, given the premises

If she has an essay to write she will study late in the library.  
 If she has some textbooks to read, she will study late in the library.  
 She does not have an essay to write.

only 4% of the subjects drew the conclusion 'She will not stay late in the library'; and given the premises

If she has an essay to write she will study late in the library.  
 If she has some textbooks to read, she will study late in the library.  
 She stays late in the library.

only 13% drew any conclusion.

The conclusion that Byrne draws from the experimental results is that

...in order to explain how people reason, we need to explain how premises of the same apparent logical form can be interpreted in quite different ways. The process of interpretation has been relatively neglected in the inferential machinery proposed by current theories based on formal rules. It plays a more central part, however, in theories based on mental models.

The first sentence could be salvaged if it is taken to mean that the interpretation process can result in assigning different logical forms to what are *prima facie* sentences with the same logical form, but the picture painted in the last sentence is rather too rosy. In any case she considers the results to be evidence against rule theories: if *modus ponens* were a mental rule, how is it possible that its application may be dependent on a premise, which is neither the major nor the minor premise of the rule?

Our conclusion from the experimental results will be rather different. To us, the pattern of results indicate that subjects sometimes use a *non-monotonic* form of logical validity, in which the addition of a premise may invalidate a conclusion previously obtained. Thus, there is no suppression of

a valid rule, *modus ponens*, but rather the application of the same reasoning scheme across the board. This scheme can be viewed both syntactically, as the application of a set of rules, and semantically, in terms of a privileged class of models. In fact, a standard completeness theorem shows that the two viewpoints are equivalent with respect to the set of inferences generated, so that no experiment based on observing inferences made by subjects can decide between the two. In any case there is no basis for the claim of the ‘mental models’ school that constructing models is somehow different from logic.

Byrne sees the main problem as explaining ‘how premises of the same apparent logical form can be interpreted in quite different ways’. The main question however is: what *is* logical form? Traditionally, it has been taken to mean the formal expression which results from translating the given sentence into a formal language; this is also how Byrne conceives of logical form. However, from a logical point of view much more is involved. Let  $\mathcal{N}$  be natural language,  $\mathcal{L}$  a formal language into which  $\mathcal{N}$  is translated. A more complete list of what is involved in logical form is given by:

1. the expression in  $\mathcal{L}$  which translates an expression in  $\mathcal{N}$
2. the semantics  $\mathcal{S}$  for  $\mathcal{L}$
3. the definition of validity of arguments  $\psi_1, \dots, \psi_n/\varphi$ , with premises  $\psi_i$  and conclusion  $\varphi$ .

The essential point is that the experimental task may be interpreted in different ways, and that this may result in different logical forms. The subjects were asked to ‘choose one of the conclusions whichever you think follows from the sentences’. ‘Follows’ can only be understood with respect to a logical form, and here there is much room for variation<sup>4</sup>. Not only has mathematical logic come up with a great many formal systems, and several interesting notions of validity (classical, constructive, nonmonotonic, dynamic), these possibilities can occur in human reasoning as well.

Take the concept of ‘validity’. Why would subjects’ understanding of ‘follows’ be the same as that of a (classical) logician? Again, this question is not meant to suggest that ‘anything goes’ in the domain of logic, but rather that ‘follows’ can mean different things to different people in different circumstances. The classical logician’s concept of ‘ $\varphi$  follows from  $\psi_1 \dots \psi_n$ ’ is

---

<sup>4</sup>Moreover, ‘follows’ need not be understood in its logical sense.

that  $\varphi$  is true in all models of  $\psi_1 \dots \psi_n$ ; thus, we must leave out of consideration all information we happen to have, beyond the premises  $\psi_1 \dots \psi_n$ . This is typically almost impossible to achieve for those without logical training<sup>5</sup>. It thus makes sense to inquire whether there might be notions of validity closer to 'common sense'.

We will now reanalyse Byrne's data with the above distinctions in mind. The outcome of the analysis will be that there is no 'suppression effect', and indeed that the observed reasoning patterns, 'fallacies' included, conform to well-known logical forms, generically known as 'common sense inference'. We will first illustrate the main logical ideas involved using a variant of nonmonotonic logic—this will explain the suppression effect and the 'denial of the antecedent' fallacy. The 'affirmation of the consequent' fallacy will then be explained by a slightly different nonmonotonic notion of validity.

In nonmonotonic logic the classical concept of validity

an argument is valid if its conclusion is true in *all* models of its premises

is changed to

an argument is valid if its conclusion is true in all *preferred* models of its premises.

The force of this of course depends on what 'preferred' is taken to mean. Typically, the preferred models are in a sense minimal; e.g. in predicate logic this would mean that they contain the minimal number of elements compatible with the premises, or that the predicates are satisfied by as few elements (or tuples) as possible. A case can be made for the claim that common sense reasoning proceeds via preferred, even minimal, models, instead of all models of premises. For example, in reasoning about events we usually care only about events which are explicitly mentioned or whose occurrence is foreseen. This is an instance of the notorious *frame problem*. If I switch on my computer, then it does not follow (classically) that the colour of the walls in my study remain the same. This is however a common sense conclusion, which can be derived formally in a theory detailing the events involved in switching on the computer and in walls changing colour. In a minimal model of this theory, the latter type of events will not occur, and so it follows (with respect to the restricted class of models) that the walls will not change colour. The term 'nonmonotonic' derives from the fact that

---

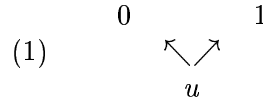
<sup>5</sup>Byrne replaced all subjects in her sample who had taken a course in logic.

the addition of a premise may render a previously valid inference invalid. This is so because what was previously a minimal model may no longer be so after the addition of a new premise.

With these ideas in place we may now introduce a specific type of non-monotonic reasoning especially relevant to conditional statements. The appropriate framework for these ideas is ‘logic programming with negation as failure’ to which we now give a very brief introduction.

**Definition 1** A (definite) clause is a formula of the form  $(\neg)p_1 \wedge \dots \wedge (\neg)p_n \rightarrow q$ , where the  $p_i$  are either propositional variables,  $\top$  or  $\perp$ <sup>6</sup>. A definite logic program is a conjunction of definite clauses.

The proper semantics for definite clauses is based on Kleene’s three-valued logic with truth values  $\{u, 0, 1\}$ . The third value  $u$  should not be thought of as a degree of truth, but as a way of expressing that the truth value is as yet undecided. The set of truth values  $\{u, 0, 1\}$  thus has the partial order  $u \leq 0$  and  $u \leq 1$ , as in the following diagram



The semantics of the connectives  $\neg, \wedge, \vee$  is then given by<sup>7</sup>

Strong Kleene operations ([64])							
		$p$	$q$	$p \wedge q$	$p$	$q$	$p \vee q$
		1	1	1	1	1	1
		0	0	0	0	0	0
$p$	$\neg p$	$u$	$u$	$u$	$u$	$u$	$u$
1	0	1	0	0	1	0	1
0	1	1	$u$	$u$	1	$u$	1
$u$	$u$	0	1	0	0	1	1
		0	$u$	0	0	$u$	$u$
		$u$	1	$u$	$u$	1	1
		$u$	0	0	$u$	0	$u$

A very important feature of the Kleene operations is that they are *monotone*, in the following sense. Let  $\varphi$  be a formula in which a proposition letter  $p$  occurs. Let  $V(p) = u$  and  $V(p) \leq V'(p)$ , so that  $V'(p)$  can be  $u, 0$  or  $1$ . Then also  $V(\varphi) \leq V'(\varphi)$ . Thus, when we obtain more information about  $p$ ,  $\varphi$  cannot change its truth value from  $0$  to  $1$ , or to  $u$ , say.

<sup>6</sup>We use  $\top$  for an arbitrary tautology, and  $\perp$  for an arbitrary contradiction

<sup>7</sup>The implication is not included in this list, because it will be provided with a non truth functional semantics.

We may consider valuations to be partial functions from the propositional language to  $\{u, 0, 1\}$ . Provide the set of three-valued valuations (i.e. models) with an ordering as follows.

**Definition 2** *Let  $V_1, V_2$  be three-valued valuations. Put  $V_1 \leq V_2$  if for all  $p$  in the domain of  $V_1$ ,  $V_1(p) \leq V_2(p)$ , where the last  $\leq$  is taken in the sense of Kleene.*

There is a canonical way to compute models for definite logic programs. The basic idea is that models (or valuations) for a definite logic program  $P$  are constructed iteratively, starting from a valuation  $V_0$  which is fully unspecified. That is, for each proposition letter  $p \neq \top$ , we put  $V_0(p) = u$ ;  $V_0(\top) = 1$ . We now proceed iteratively. Suppose  $V_n$  has been constructed; we then construct  $V_{n+1}$  such that  $V_n \leq V_{n+1}$ . Put  $V_{n+1}(p) = 1$  if and only if for *some* clause  $\varphi \rightarrow p$  in  $P$ :  $V_n(\varphi) = 1$ . Put  $V_{n+1}(p) = 0$  if and only if for *all* clauses  $\varphi \rightarrow p$ :  $V_n(\varphi) = 0$ . Otherwise, put  $V_{n+1}(p) = u$ <sup>8</sup>. This construction embodies the notion of ‘negation as failure’: if a proposition cannot be shown to be true, it must be false. We will have occasion to modify this notion later, but it embodies an important intuition. We contend that real-life conditionals can be captured much more adequately by logic program clauses of the form  $p \wedge \neg ab \rightarrow q$ , where  $ab$  is a proposition letter indicating that something abnormal is the case. Negation as failure now ensures that if  $p$  and there is no positive evidence for  $ab$ , then we may conclude  $q$ . We will thus want to keep negation as failure as applied to the special class of abnormality propositions, even when we reject it for all propositions, as we will do below. Negation as failure gives rise to nonmonotonicity, because our happy certainty that nothing abnormal is going on may be shattered by the addition of a further premise.

One can check, using the monotonicity of the Kleene operations, that  $V_n \leq V_{n+1}$ . A little thinking now shows that from some  $n$  onward,  $V_n = V_k$ ,  $k > n$ . If  $n$  is the smallest integer for which this happens, we will call  $V_n$  the *minimal model*, minimal in the sense that as few propositions are true as is compatible with the data. In general the minimal model leaves the truth values of some proposition letters undecided, although this happens only for ‘pathological’ clauses such as  $\neg p \rightarrow p$ . In the more usual case, where such circularities do not occur, all propositions will be decided. However, this does not detract from the necessity to use Kleene’s three-valued logic, which makes the iterative construction possible.

---

<sup>8</sup>Note that we need the Kleene truth tables for  $\neg, \wedge, \vee$  here, but no truth table for  $\rightarrow$ : the conditional is defined ‘operationally’ by means of the preceding construction.

Now comes the main hypothesis: *when making inferences, people do not consider all models of the premises, but only the minimal model.* The question is therefore: what is true on minimal models? A general result says that what is true on the minimal model is captured by the *completion* of the program.

**Definition 3** *The completion of a definite logic program is defined by means of the following procedure:*

1. *if a propositional variable  $p$  does not occur in the consequent of a clause, add a formula  $\perp \rightarrow p$ .*
2. *if a formula is of the form  $q$ , i.e. the consequent of a clause with empty antecedent, add a formula  $\top \rightarrow q$ .*
3. *for each propositional variable  $q$ , collect the clauses  $\varphi_i \rightarrow q$  with  $q$  as consequent, form  $\bigvee_i \varphi_i$  and add the formula  $q \leftrightarrow \bigvee_i \varphi_i$ .*

Here,  $\leftrightarrow$  is semantically interpreted by  $V(\psi \leftrightarrow \varphi) = 1$  iff  $V(\psi) = V(\varphi)$ , and  $= 0$  otherwise.

The general result alluded earlier can then be stated as

**Theorem 1** *Let  $P$  be a definite logic program,  $\text{comp}(P)$  its completion. Then  $V'(\text{comp}(P)) = 1$  if and only if for the minimal model  $V$  of  $P$ :  $V \leq V'$ .*

With these technical details in place, we are now ready to show that the reasoning pattern exhibited by Byrne's subjects can be explained if we assume that these subjects have a nonmonotonic competence model. We will proceed in two stages. At first we apply the construction outlined above, including negation as failure. This will explain some, but not all, of Byrne's data. In order to capture the remainder of the data, we take our cue from a paper by Dieussaert et al. [22], which suggests that there are at least two semantics for program clauses (i.e. conditionals) used by subjects.

We will represent the conditionals in Byrne's experiment as *definite clauses* of the form  $p \wedge \neg ab \rightarrow q$ , where  $ab$  is a proposition which indicates that something abnormal is the case, i.e. a possibly disabling condition. It is essential that the conditionals are represented as being part of a definite logic program, so that their semantics is operational not truth functional. We show that on the basis of this interpretation, all inferences except 'affirming the consequent' correspond to valid argument patterns. The latter inference type necessitates a further refinement of the notion of validity, and will be treated separately.

**Modus ponens for a single conditional premise** Suppose we are given a conditional  $p \wedge \neg ab \rightarrow q$  and the further information that  $p$ . By negation as failure we may conclude that on the minimal model  $\neg ab$ , so that  $q$  holds there as well.

**A ‘fallacy’: denial of the antecedent** Suppose we are given a conditional  $p \wedge \neg ab \rightarrow q$  and the further information  $\neg p$  (i.e.  $p \rightarrow \perp$ ). At first sight it may seem that a formula of the form  $p \rightarrow \perp$  cannot be part of a program. However, as long as  $p$  does not occur in the head of a clause there is no problem, because it would be assigned the value 0 anyway<sup>9</sup>. Since the set up does not contain any further information about  $ab$ , by theorem 1 the minimal model of this definite program satisfies  $p \leftrightarrow \perp$ ,  $ab \leftrightarrow \perp$  and  $p \leftrightarrow q$ , whence  $\neg q$  is true on the minimal model.

**Additional premises** As we have seen, if the scenario is such that nothing is said about  $ab$ , minimisation sets  $ab$  equal to  $\perp$  and the conditional  $p \wedge \neg ab \rightarrow q$  reduces to  $p \rightarrow q$ . Now suppose that the possibility of an abnormality is made salient by adding a premise ‘if the library is open, she will study late in the library’. This adds formulas  $r \rightarrow q$  (but see below), *but also*  $\neg r \rightarrow ab$ . Minimising  $ab$  makes it equivalent to  $\neg r$ , so that the first conditional becomes  $p \wedge r \rightarrow q$ . Actually, the situation is symmetric, in that the first conditional highlights a possible abnormality relating to the second conditional. The circumstance that the library is open is not sufficient incentive to go study there, one must have a purpose for doing so. The second conditional accordingly becomes  $r \wedge \neg ab' \rightarrow q$  with defining condition  $\neg p \rightarrow ab'$  for  $ab'$ . Minimisation yields  $\neg ab' \leftrightarrow p$ , so that after minimisation the first and second conditionals both collapse into the conditional  $p \wedge r \rightarrow q$ . Now *modus ponens* is suppressed, because the necessary premise  $r$  is lacking. Thus we see nonmonotonicity at work: the minimal model for the case of an additional premise is essentially different from the minimal model of a single conditional premise plus factual information.

If we were to add the premise  $\neg p$  instead of  $p$ , we would have  $\neg(p \wedge r)$ , whence  $\neg q$  is true in the minimal model, so that ‘denying the antecedent’ will not be suppressed, as observed.

**Alternative premises** Again we have two conditional premises which must be formalised as  $p \wedge \neg ab \rightarrow q$  and  $r \wedge \neg ab' \rightarrow q$ . But in this case the

---

<sup>9</sup>From this perspective, *modus tollens* arguments thus belong to a different category!

alternatives do not make salient possible obstacles, so that after minimisation both  $ab$  and  $ab'$  are set equal to  $\perp$ . In the case of the *modus ponens* inference, the completion then becomes  $(p \vee r \leftrightarrow q) \wedge p \wedge (r \rightarrow \perp)$ , hence  $q$  is true in the minimal model. In the case of ‘denial of the antecedent’, however, the present set up gives the wrong prediction. The completion becomes  $(p \vee r \leftrightarrow q) \wedge (p \rightarrow \perp) \wedge (r \rightarrow \perp)$ , whence  $\neg q$  is true in the minimal model. As we have seen, in Byrne’s study this answer was given by only 4% of the participants. However, in Dieussaert et al.’s study [22] 22% gave this answer, so that it makes sense to look at a definition of validity which sanctions this answer.

**A modified construction** In the above we outlined a standard logic programming approach to conditionals, standard in the sense that it embodies negation as failure. As we have seen, full negation as failure may be a bit too strong; in the case of alternative premises, it yields DA where the majority of subjects is not inclined to make this inference. The following modified iterative construction of a minimal model comes closer to the observations.

Let  $V_0$  be as before. The step from  $V_n$  to  $V_{n+1}$  now proceeds as follows. Put  $V_{n+1}(p) = 1$  if and only if for some  $\varphi: \varphi \rightarrow p \in P$  and  $V_n(\varphi) = 1$ . If  $p$  not of the form  $ab$ , put  $V_{n+1}(p) = 0$  if and only if there exists a  $\varphi$  such that  $\varphi \rightarrow p \in P$ , and for all  $\varphi$  such that  $\varphi \rightarrow p \in P: V_n(\varphi) = 0$ . Lastly, if  $ab$  does not occur as consequent of a clause, put  $V_{n+1}(ab) = 0$ .

The corresponding notion of completion is obtained by forming the equivalence  $q \leftrightarrow \bigvee_i \varphi_i$  only if there exists at least one  $\varphi$  such that  $\varphi \rightarrow q \in P$ , and by applying negation as failure only to proposition letters of the form  $ab$ .

The difference with the preceding construction is thus that, except for abnormality propositions, the truth values of proposition letters not occurring as the consequent of a clause are not decided in advance. As a consequence, we do not have DA in the case of an alternative premise. For if our data are

$$\begin{array}{l} p \wedge \neg ab \rightarrow q \\ r \wedge \neg ab' \rightarrow q \\ \neg p \end{array}$$

then there will be a model in which  $r$  has value  $u$ , so that nothing follows about  $q$ . Apart from this, we obtain the same results as with the original construction: in the case of a single conditional premise MP and DA, in the case of an additional premise DA and suppression of MP, and in the case of an alternative premise MP.

**Strategies** So far we have adopted what Dieussaert et al. [22] call an ‘integration strategy’: the premises, both conditionals and facts, are taken jointly when constructing a minimal model, and it is assumed that no more premises will be supplied. This justifies the condition in the model construction which assigns 0 to a proposition letter  $ab$  that does not occur in the head of a definite clause. For if no more premises can be added, there is no way in which  $ab$  can be made true, so we may assume it is false (the *closed world assumption*). The integration strategy as outlined predicts that AC will not occur. For if the program  $P$  consists of  $p \rightarrow q$ ,  $q$ , then, using the modified construction, after completion we obtain  $(p \vee \top) \leftrightarrow q$ , so that nothing is predicted about the truth value of  $p$ .

The strategy just outlined is not the only possibility: Dieussaert et al. [22] obtained some evidence for an ‘amendment strategy’ where premises are considered one at a time, and the minimal model is updated continuously. The name ‘amendment’ derives from Dieussaert et al.’s view that a conclusion based on the first conditional and the categorical premise is amended in the light of the second conditional. We prefer the name ‘update strategy’, since this is less committed to what is actually updated. In fact we believe that already in the case of a single conditional premise some updating may be going on, and that AC is evidence for this.

In any case it is not necessary to view this as a processing strategy only. It can also be seen as yet another notion of validity; in fact, it is a notion of validity underlying various forms of dynamic logic. The closed world assumption is now no longer in force, and therefore the construction of models has to be adapted. This strategy will yield answers in several cases where the integration strategy did not provide an answer, and we will also be able to treat instances of ‘backward’ reasoning, namely AC and MT.

A difference with the preceding construction is that the models are no longer constructed iteratively, but by means of a largely indeterministic procedure. Given a set of premises, first separate the facts from the rules, and collect the latter into a definite logic program  $P$  consisting of definite clauses of the form  $\varphi \rightarrow q$  with  $\varphi$  nontrivial. Form a completion  $comp(P)$  of this program, not using negation as failure. The completion will thus consist of a number of statements of the form  $q \leftrightarrow \bigvee_i \varphi_i$ . The meaning of  $comp(P)$  is that we remain agnostic about atomic facts in the absence of further information, but apply a kind of closed world assumption to the ‘laws’ occurring in  $P$ . If  $P$  says that  $q$  can be a consequence of  $\varphi_1, \dots, \varphi_n$ , then we assume that these are all the ways in which  $q$  can come about.

In the second stage we may now add atoms or negations of atoms; let us call the set of added formulas *Fact*. We furthermore add  $\neg ab$  for all those  $ab$

which do not occur in the consequent of a clause; let us call this set  $NegAb$ . We are now looking for models of  $comp!(P)+Fact+NegAb$ , assuming this set is consistent. However, it is no longer clear how to construct ‘minimal’ models, since the possible presence of negative facts disables the iterative construction (‘forward reasoning’) sketched earlier<sup>10</sup>. We are thus left with a tableau-like procedure, which will in general yield several possible models.

We should emphasise that the ‘update strategy’ really is another notion of *validity*. The class of models considered in the definition of validity is now restricted to the models which can be obtained in the way outlined above. That is, we do not look at all (classical) models of  $P+Fact$ , but focus instead on the more restricted set of models for  $comp!(P)+Fact+NegAb$ . This class of models differs from the minimal models corresponding to the integration strategy, but it is also not equal to the set of all models of the premises. This notion of validity is again nonmonotonic. For example, a conditional premise with  $ab$  as consequent may be added, overturning a previous judgment  $\neg ab$ . More importantly, a conditional premise  $\varphi \rightarrow q$  may be added to  $P$  when this program already features a clause with  $q$  as consequent. This will in general increase the number of models, and may make previously valid inferences invalid, as will become clear below.

**Example: AC for a single conditional premise** Suppose we have a single conditional premise  $p \rightarrow q$  and a fact  $q$ . As we have seen, the integration strategy would yield the completion  $(p \vee \top) \leftrightarrow q$ , from which, barring negation as failure, nothing can be concluded about  $p$ . By contrast, the update strategy first forms  $p \leftrightarrow q$ , and updates the truth value of  $p$  in the light of incoming information about  $q$ . This gives both AC and MT. We also see that the model constructed by the update strategy does not coincide with the set of classical models of the premises, for classically  $V(p) = 0, V(q) = 1$  is also a model.

**The update strategy for several conditional premises** In the case of an additional premise,  $P$  consists of  $p \wedge \neg ab \rightarrow q, r \wedge \neg ab' \rightarrow q, p \rightarrow ab'$ , and  $r \rightarrow ab$ . After completion we obtain as before  $p \wedge r \leftrightarrow q$ . Adding  $q$  as a fact leads to the conclusion  $p$ ; by contrast, adding  $\neg q$  does not determine a unique conclusion. Of course, adding  $p$  still does not yield  $q$ . This pattern squares with the empirical results.

In the case of an alternative premise, the minimal model will now only satisfy  $\neg ab \wedge (p \vee r) \leftrightarrow q$ , with all of  $p, q, rab$  undecided. In this situation,

---

<sup>10</sup>In fact, we claim that this explains part of the difficulty of *modus tollens*.

$ab$  is set to false, and after doing this we obtain  $p \vee r \leftrightarrow q$ . Adding  $q$  does not yield a unique conclusion, but both  $p$  and  $\neg q$  do. Again this squares with the observed results.

**Suppression suppressed?** Dieussaert et al. proposed the distinction between the two strategies in order to account for their finding that by enlarging the set of possible answers, the suppression effect for valid inferences is mitigated and the fallacies are suppressed. They claim that Byrne's experiment is flawed in that *only for an atomic proposition or its negation* may a subject judge whether it follows from the premises supplied. This would make it impossible for the subject that draw a conclusion which pertains to both conditional premises. Accordingly, they also allow answers of the form  $(\neg)A(\wedge)(\vee)(\neg)B$ , where  $A, B$  are atomic. Unfortunately, since they also require subjects to choose only one answer among all the possibilities given, the design is flawed. This is because there exist dependencies among the answers (consider e.g. the set  $p \vee q, p \wedge q, p, q$ ) and some answers are always true (e.g.  $p \vee \neg p$ ). Thus, the statistics yielded by the experiment are not interpretable. Nevertheless, the reasoning behind the experiment is interesting. For example, they argue that, in the case of the argument

Is she has an essay to write, then she will study late in the library.

If she has some textbooks to read, she will study late in the library.

She does not have an essay to write.

the integration strategy and the update strategy will yield different answers, and that this difference can only be uncovered by means of an enlarged answer set.

We have seen that the modified integration strategy indeed does not assign a truth value to

She will study late in the library.

According to Dieussaert et al., the update strategy could instead yield the answer 'Either she will not study late in the library, or she has a textbook to read', with the following reasoning. The first conditional premise and the categorical premise are first taken into consideration, and a (putative) conclusion is formulated on this basis. Closed world reasoning thus gives 'She will not study late in the library.' The second step then considers the other conditional premise, and determines whether the conclusion obtained in the

first step must be updated. Now the possibility ‘She has some textbooks to read’ is also considered. Thus, the world turns out to be a little less closed, but not completely open either, and we are forced to a conditional conclusion: ‘If she studies late in the library, she has some textbooks to read’. In fact, 67% of the subjects gave this answer (albeit in disjunctive form stated above, since only this form featured in the answer set). Again, the statistics of the experiments are vitiated by the flaw in the design of the answer set, but the argument that integration strategy and update strategy should yield different answers is sound.

We will now reproduce this reasoning in our more formal setting, using the update strategy. Our explanation is slightly different, and is based on the assumption that factual information is processed *after* the conditionals have been taken care of. Consider the case of two conditional premises with alternative antecedents, and the inference ‘denial of the antecedent’:

$$\begin{array}{l} p \wedge \neg ab \rightarrow q \\ r \wedge \neg ab' \rightarrow q \\ \neg p. \end{array}$$

We claim that first the two conditional premises are considered, yielding a model where  $q \leftrightarrow (p \wedge \neg ab) \vee (r \wedge \neg ab')$  is true, and all proposition letters are undecided. We then update the model with the further information  $\neg p$ . Assuming  $\neg p$  represents all the factual information, both  $ab$  and  $ab'$  can be set to false, which yields  $q \leftrightarrow (p \vee r)$ . It follows that if  $q$  is true, so is  $r$ .

Dieussaert et al. interpret this frequently given answer as evidence that Byrne’s view of the suppression effect needs to be qualified: given a richer set of possible answers, there would no longer be an unambiguous suppression effect. Consider the paradigm case of MP with an additional premise. Here, 69% chose (an answer implying)  $q$ , which can be interpreted as suppression when compared to the usual rate for MP, but 18% now selected the answer ‘if  $r$  then  $q$ ’ (again, given in disjunctive form). The reader may check that the update strategy would also yield this answer. Observe that, although ‘if  $r$  then  $q$ ’ looks like the simple repetition of the second premise, it probably is a material implication (cf. the disjunctive formulation), whereas the second premise is not. This would explain why subjects regard their answer as informative. It will be clear from the above that we do not believe in a suppression effect, but the data obtained by Dieussaert et al. are useful in showing that at least two different forms of nonmonotonic reasoning are involved.

**Conclusion** Let us retrace our steps. Byrne claimed that both valid and invalid inferences can be suppressed, based on the *content* of supplementary material; therefore, the *form* of sentences would determine only partially the consequences that people draw from them. Our analysis is very different. Consider first the matter of form and content. We believe that logical form is not simply read off from the syntactic structure of the sentences involved, but is assigned on the basis of 'content'—not only that of the sentences themselves, but also that of the context. In this case the implied context—a real-life situation of going to a library—makes it probable that the conditionals are not material implications but some kind of defaults. We then translate the conditionals, in conformity with this meaning, into a formal language containing the  $ab$ ,  $ab'$ , ... formulas. However, translation is just the first step in imposing logical form. The second step consists in associating a semantics and a definition of validity to the formal language. For the semantics we have chosen three-valued Kleene logic, and the definition of validity is given in terms of (several notions of) minimal models. Once logical form is thus fixed (but not before!), one may inquire what follows from the premises provided. In this case the inferences observed in the majority of Byrne's subjects correspond to valid inferences (given the assignment of logical form). Hence we would not say that content has beaten form here, but rather that content determines logical form. There are many different ways of assigning logical form, and that of classical logic is not by any means the most plausible candidate for common sense reasoning. By sticking to classical logic as *the* competence model, the psychology of reasoning has been forced to study spurious phenomena.

It is one thing to claim, as we have done, that Byrne's subjects have on the whole responded normatively, but it is a further step to argue that their actual processing also follows the lines indicated above. The inference procedure sketched above is quite intricate, much more so than reasoning in plain classical propositional logic. On the other hand, we contend that subjects would also have much more occasion to reason according to this 'common sense' notion, so that perhaps classical logic is the more difficult abstraction.

[to be added: section on 'jumping to conclusions']

### 3.3 'illusory inferences'

An allegedly stronger form of 'nonlogical' reasoning can be found in work on 'reasoning illusions' by Johnson-Laird and colleagues. (Johnson-Laird

& Savary [60]; Johnson-Laird, Legrenzi, & Girotto [61]; Johnson-Laird & Byrne [59]). Their interest in these problems is that mental models theory assumes that subjects ‘only represent explicitly what is true’, and that this gives rise to ‘illusory inferences’. The following material was presented with the preface that both statements are about a hand of cards, and one is true and one is false:

1. If there is a king in the hand, then there is an ace.
2. If there is *not* a king in the hand, then there is an ace.

Select a conclusion:

There an ace in the hand

There is not an ace in the hand

There may or may not be an ace in the hand.

Johnson-Laird & Savary ([60] report that 15 out of 20 subjects concluded that that there *is* an ace in the hand, and the other five concluded that there might or might not be an ace in the hand. They claim that the correct answer is that there is not an ace in the hand.

What is claimed to be novel here is not subjects’ supposedly errorful reasoning (which is ubiquitous in this literature), but in subjects’ refusal to give up their supposedly erroneous inferences when they are pointed out. This is likened to the persistence of visual illusions even when subjects know that their perceptual judgements are wrong. One might question the analogy, since subjects who do not accept the inferences do not experience a conflict between what they ‘see’ and what they believe. Nevertheless, it is of some interest to compare their results with those from our two-rule task.

There are many problems with Johnson-Laird & Savary’s claims. They rely on a very particular interpretation of the conditionals to get their conclusion, namely a material interpretation applying to a single card (not implicitly quantified over a domain of cards). Yet we know that naive subjects have the greatest difficulty making this interpretation, and that there are other more reasonable interpretations of the conditionals presented. A much more reasonable assumption is that subjects adopt an interpretation on a domain of indefinitely many hands of cards, to the effect that there is a regularity that if there is a king, then there is an ace, and similarly for the other conditional (exactly the expected interpretation in our two rule task). On this interpretation the correct answer is that there might or might not be an ace in the hand. Exactly the same proportion of subjects got the task correct on this interpretation as got our two-rule task correct.

An interesting question about Johnson-Laird & Savary's observations is: to why so many subjects make the stronger conclusion that there positively must be an ace in the hand. Here there are several possibilities, of which perhaps the most likely is that the subjects think (not unreasonably) in terms of one of the rules *applying* and the other not, and they confuse (not surprisingly) the semantics of applicability with the semantics of truth (just as our tutoring dialogues reveal our subjects frequently did). This is exactly the semantics familiar from the *IF ... THEN ... ELSE* construct of imperative computer languages.

Indeed, Johnson-Laird, Legrenzi, & Girotto ([61] have a version of this experiment with slightly changed wording, involving an exclusive disjunction:

If there is a king in the hand, then there is an ace, or else if there is not a king in the hand, then there is an ace.

A subject trying to parse this sentence face a scoping problem: does 'or else' scope over the entire conditional or only over its antecedent? The later reading is strongly suggested by the *IF ... THEN ... ELSE* construct; one then gets something like

If there is a king in the hand, then there is an ace, or else, *i.e.*  
if there is not a king in the hand, then there is an ace.

If one clause applies and the other doesn't, then it obviously follows that there is an ace. Whether the alternativeness of the rules is expressed metalinguistically (by saying one is false and one true) or object-linguistically (with an exclusive disjunction), thinking in terms of applicability rather than truth is a great deal more natural and has the consequence observed. If the subjects are thinking in terms of applicability, then this is another example of the complexities of semantic relations in these tasks which is emphasised in these notes.

We will put our money on the subjects having the more plausible interpretation of the conditionals here and the experimenters suffering an illusion of an illusion. The most persistent and damaging aspect of this illusion is that logic proposes only one interpretation for conditionals.



## Chapter 4

# Propositional logic—the hard cases

Whereas the ‘illusory inferences’ discussed previously can be argued to be relatively innocuous and mainly due to a tin ear for language on the part of the experimenters, there do exist cases where it is well-nigh impossible to elicit reasoning according to classical logic. The most famous of these is

### 4.1 Wason’s 4 card task

In 1968, Peter Wason proposed a now classic experiment which combines reasoning with an action to be performed, namely selection of a subset of a set of four cards (for this reason the 4 card task is often called the selection task). Here is an example of such a task:

Below is depicted a set of four cards, of which you can see only the exposed face but not the hidden back. On each card, there is a number on one of its sides and a letter on the other.

Also below there is a rule which applies only to the four cards. Your task is to decide which of these four cards you *must* turn in order to decide if the rule is true. Don’t turn unnecessary cards. Tick the cards you want to turn.

**Rule:** *If there is a vowel on one side, then there is an even number on the other side.*

**Cards:**



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 [113]).

This experiment has been repeated with many different populations and with many different designs, and with few exceptions (to be discussed below) the same distribution of answers was observed.

#### 4.1.1 Deontic variants

The psychologists involved in these experiments have often seen these observations as knock-down arguments against the employment of formal theories in explaining students' behaviour (e.g. Johnson-Laird and Wason [114]; Johnson-Laird and Byrne [57]). This response has especially been engendered by what are known as *thematic* or *content* effects. Early after Wason's initial experiment, Wason and Shapiro [115]) and Wason and Johnson-Laird [114] 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. Here is an example of such a 'thematic condition':

Imagine you are in a bar. Below is depicted a set of four cards, of which you can see only the exposed face but not the hidden back, and each card corresponds to a drinker in your imagined bar. A drink is named on one side of each card, and its drinker's age on the other.

On the wall of the bar appears a large notice reminding customers of the drinking age law that applies in the bar:

**To drink whisky you must be over 18 years of age**

Your task is to decide which cards (if any) you *must* turn in order to decide whether any of the drinkers represented by the cards is breaking the law. Don't turn unnecessary cards. Tick the cards you want to turn.

**Cards:**

whiskey	orange juice	19	17
---------	--------------	----	----

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* [102] 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 [71]). 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 [39] showed that regulations provided particularly facilitating kinds of thematic linkage. Cheng and Holyoak [15] 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 [19]), 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 [33]). 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 [112] 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 elicited 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. [88] 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 [38] earlier failed to find such facilitation. In fact, 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 material which might best be described as incorporating qualitative laws of physics and obtained good performance from their student subjects.

### 4.1.2 Tutorial experiments

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 loses 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 leaves no room for such answers. 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 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 suggested a range of more conventional experiments, the results of which will be described later.

The sessions were videotaped and then transcribed. The experiment was performed in two runs, March 1999 (19 subjects) and July 1999 (10 subjects). 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 which might, or might not, be 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'. This condition will be referred to as the *paraphrase task*.

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

The first example shows that 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 [Standard Wason task] *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 A on the back of 7, and notes that this would mean the rule was false of the card, but then declines to choose the card (or revise an earlier failure to choose it).

*Example* Subject 3 *E*. OK Lastly the 7.

*S.* Well I wouldn't pick it.

*E.* But what would it mean if you did?

*S.* Well, if there is an A then that would make the rule false, and if there was a K, 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 7/A and 7/K 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.

### 4.1.3 Standard explanations, and why they fail

E.g. inability to perform *modus tollens* inferences, verification bias, . . .

#### Matching bias: the 'no-processing explanation'

Evans (see for example the review in Evans, Newstead and Byrne [24]) defines 'matching strategy' as the choice of cards which match the atomic parts 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 1] *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 [76] 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.

#### 4.1.4 Verifying and falsifying

Wason’s initial explanation of his findings, which he took to be an application of Popper’s claims in the philosophy of science, is that subjects would be tempted to turn the  $q$  card, because  $q/p$  confirms the rule, whereas  $q/\neg p$  is irrelevant. Before we discuss this in detail, let us give an illustration.

Subject 3 in a standard Wason task 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.

Subject 13 [Standard Wason task] *S.* (Turns the 7) ... 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 ...<sup>1</sup>

---

<sup>1</sup>For another striking example, cf. subject 26 in section 5.1.

The first thing to be said is that there is a terminological issue about *verification*. 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, and then stick to the term. 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 transcripts 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.

Subject 1. [Standard Wason task] *E*. What could there be on the back of the 4?  
 [Subject writes K and A.]  
*E*. OK. And in the case of the A?  
*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 A on the back of the 4?  
*S*. Doesn't matter, doesn't either make it true or false.  
*E*. OK. So do you want to turn over the 4?  
*S*. Not particularly.

The next example presents subject 4 who gave the normative response in one version of the Wason task and now struggles with a slight variant.

Subject 4.  
*E*. You picked the A, would you pick anything else?  
*S*. Yes I would [...] the K is irrelevant. OK the 7? [...] If there's a K there then that's fine, but if there's an A then that falsifies. [...] This one is a 4, it could be an A or there could be a K, so yes I would turn it over, if there was a K then that would falsify the rule. The K doesn't come into it because if there's a 4 - it doesn't say if there's a 4 there has to be a K. Does it ... shit ...  
*E*. Turn over the cards you want to turn.

*S.* [picks up the 4] Well this falsifies the rule because ... no shit, does it? ... Yes it does because there isn't an A and 4 combination.

*E.* Turn over the other cards now.

*S.* This is a 7. I have to turn it over to check whether there is an A. If there's an A then that also falsifies the rule. Oh and there is. I don't want to touch the K. Now I'm going to turn the A to see if there is a 4 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.

Subject 6.[Standard Wason task]

*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 ( $\geq 40\%$ ), 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 3 and in one experiment 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.

## 4.2 2-rule selection task

Stenning and van Lambalgen [98] introduced a novel task comprising two rules, where subjects were instructed that one is true and the other false, and are asked to decide which is which. The rules are

1. if there is a U on one side, then there is an 8 on the other side
2. if there is an I on one side, then there is an 8 on the other side

given the background rule that one side contains U or I, and the other side contains 3 or 8. In the tutorial version of this experiment, subjects were presented with real cards lying in front of them on the table. The cards shown were UI83. 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. In this case, both U and I carried an 8, 8 carried an I, and 3 a U.

The motivation for introducing this manipulation was twofold. First, the Bayesian approach due to Oaksford and Chater (cf. sectionBayes) postulates that in solving the standard Wason task, subjects always compare the rule given to the unstated alternative hypothesis that antecedent and consequent are independent. We were thus interested in seeing what would happen if subjects were presented with explicit alternatives. Second, we had a hunch that explicitly telling the subject that one rule is true and one false, would background a number of issues concerned with the notion of truth, such as the possibility of exceptions. Whether that is actually so is a moot point, but the experimental manipulation turned out to be unexpectedly fruitful; while struggling through the task, subjects often gave clear indications of where their difficulties lay.<sup>2</sup> Below we give excerpts from the tutorial dialogues which highlight these difficulties. Precisely because

---

<sup>2</sup>In that sample the baseline testing prior to tutoring showed no enhancement of performance, although tutoring in the two-rule task was more effective than in the classical task.

many semantic difficulties come to the surface in this task, it might lead to increased performance, and so it appears to be a good experiment to repeat in a standard classroom format.

### 4.3 Subjects' understanding of truth and falsity

#### 4.3.1 The logic of 'true'

On a classical<sup>3</sup> understanding of the two-rule task, the competence answer is to turn the 3; this would show which one of the rules is false, hence classically also which one is true. This classical understanding should be enforced by explicitly telling the subjects that one rule is true and the other one false. Interestingly, some subjects refuse to be moved by this instruction, insisting that 'not-false' is not the same as 'true'. Since it is ultimately the notion of truth that determines the logic, these subjects are thus guided by some nonclassical logic. This logic could be one of a family of three-valued logics, where 'not-false' includes the possibility 'undecided', or it could be of the modal variety, where, very roughly speaking,  $A \rightarrow B$  is true if there is a necessary connection (e.g. a proof) linking  $A$  and  $B$ <sup>4</sup> or it could be generated by the understanding of 'false' discussed below (section 4.3.1).

*Subject 17.*

*S.* [Writes miniature truth tables under the cards.]

*E.* OK so if you found an I under the 3, you put a question mark for rule 1, and rule 2 is false; if you turned the 3 and found a U, then rule 1 is false and rule 2 is a question mark. So you want to turn 3 or not?

*S.* No.

*E.* Let's actually try doing it. Turn over the U, you find a 3, which rule is true and which rule is false?

*S.* (Long pause)

*E.* Are we non the wiser?

*S.* No, there's a question mark.

---

<sup>3</sup>We use 'classical' as shorthand for 'according to classical logic', as opposed to, say, intuitionistic logic. Although the psychology of reasoning typically assumes that the competence model is classical logic, this section will show that alternative models should not be ruled out *a priori*.

<sup>4</sup>Examples are intuitionistic and relevance logic. Some subjects, when reading the rule(s) aloud actually inserted a modality:

*Subject 13.* [Standard Wason task]

*S.* ...if there is an A, then there is a 4, necessarily the 4...[somewhat later]...if there is an A on one side, necessarily a 4 on the other side...

*E.* It could have helped us, but it didn't help us?

*S.* Yes.

*E.* OK and the 3.

*S.* Well if there is a U then that one is disproved [pointing to the first rule] and if there is an I then that one is disproved [pointing to the second rule]. But neither rule can be proved by 3.

*E.* Turn over the last card [3] and see what's on the back of it... so it's a U. What does that tell us about the rule?

*S.* That rule one is false and it doesn't tell us anything about rule 2?

*E.* Can't you tell anything about rule 2?

*S.* No.

The subject thinks falsifying rule 1 does not suffice and now looks for additional evidence to support rule 2. In the end she chooses the 8 card for this purpose, which is of course not the competence answer even when 'not-false' is not equated with 'true' (the I card would have to be chosen). Here are two more examples of the same phenomenon.

*Subject 8.*

*S.* I wouldn't look at this one [3] because it wouldn't give me appropriate information about the rules; it would only tell me if those rules are wrong, and I am being asked which of those rules is the correct one. Does that make sense?

*Subject 5.*

*E.* What about if there was a 3?

*S.* A 3 on the other side of that one [U]. Then this [rule 1] isn't true.

*E.* It doesn't say...?

*S.* It doesn't say anything about this one [rule 2].

*E.* And the I?

*S.* If there is a 3, then this one [rule 2] isn't true, and it doesn't say anything about that one [rule 1].

The same problem is of course present in the standard Wason task as well, albeit in a less explicit form. If the cards are AK47, then turning A and 7 suffices to verify that the rule is not false; but the subject may well wonder whether it is therefore true. Let us already note here that it is precisely this difficulty which is absent in the case of deontic rules such as

If you want to drink alcohol on these premises, you have to be over 18.

Such a rule cannot be shown to be true; at most we can establish that it is not violated. So in the deontic case, subjects only have to do what they find easy in any case.

### The logic of ‘false’

Interesting things happen when one asks subjects to meditate on what it could mean for a conditional to be false. As indicated above, the logic of ‘true’ need not determine the logic of ‘false’ completely. The paraphrase task alluded to above showed that a conditional ( $p \rightarrow q$ ) being false, i.e. is often (*i.* 50%) interpreted as  $p \rightarrow \text{not } q$ ! (We will refer to this property as *strong falsity*.) This observation is not ours alone: Fillenbaum observed that in 60% of the cases the negation of a causal temporal conditional  $p \rightarrow q$  (‘if he goes to Amsterdam, he will get stoned’) is taken to be  $p \rightarrow \text{not } q$ ; for contingent universals (such as the rule in the selection task) the proportion is 30%. In our experiment the latter proportion is even higher. Here is an example of a subject using strong falsity when asked to imagine what could be on the other side of a card.

Subject 26 [Standard Wason task; subject has chosen strong falsity in paraphrase task]

*E.* So you’re saying that if the statement is true, then the number [on the back of A] will be 4. ... What would happen if the statement were false?

*S.* Then it would be a number other than 4.

Subject 18 [subject has chosen strong negation, has selected A and 4, thinks K is irrelevant] *E.* And the 7?

*S.* It could have an A yes, but if that rule is true it will have another letter.

*E.* And the 4?

*S.* The 4 should have an A 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).<sup>5</sup> In our case however, although in the baseline task  $p$  was chosen as frequently as

---

<sup>5</sup>It has sometimes been suggested (e.g. in Johnson-Laird and Byrne [57], p. 66) that strong negation is a consequence of taking the antecedent as a *presupposition*. This is analogous to Haiman’s argument ([41], 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.

$p, q$ , this response became rare after tutoring for the right interpretation of the anaphora (see section 4.6), suggesting that the  $p$  response is due rather to constant anaphora.

Note that strong falsity encapsulates a concept of necessary connection between antecedent and consequent in the sense that even counterexamples are no mere accidents, but are governed by a rule. If a subject believes that true and false in this situation are exhaustive, this could reflect a conviction that the cards have been laid out according to *some* rule. It is interesting to see what this interpretation means for card choices in the selection tasks. If a subject has strong negation but still believes true and false are exhaustive, then (in the standard Wason task) *either* of the cards  $p, q$  can show that  $p \rightarrow q$  is not-false, hence true. Unfortunately, in the standard set up 'either of A, 4' is not a possible choice. In the tutorial experiment involving the two-rule task subjects were at liberty to make such choices. In this case strong falsity has the effect of turning each of the two rules into a biconditional, 'U if and only if 8' and 'I if and only if 8' respectively. *Any* card now distinguishes between the two rules, and we do indeed find subjects emphatically making this choice:

*E.* OK so you want to revise your choice or do you want to stick with the 8?

*S.* No no ... I might turn all of them.

*E.* You want to turn all of them?

*S.* No no no just one of them, any of them.

Perhaps the customary choice of  $p, q$  in the standard task is the projection of 'either of  $p, q$ ' onto the given possibilities. Another option is that some subjects have a biconditional reading of 'if...then' together with strong falsity; in this case both  $p$  and  $q$  are necessary. These considerations just serve to highlight the possibility that a given choice of cards is made for very different reasons by different subjects, so that by itself statistical information on the different card choices in the standard task may be of little significance.

### **Truth of the rule and 'truth of the card'**

Subjects are persistently confused about several notions of truth that could possibly be involved. The intended interpretation is that the domain of discourse consists of the four cards shown, and that the truth value of the rule is to be determined with respect to that domain. This interpretation is however remarkably difficult to get at. An alternative interpretation is that the domain is some indefinitely large population of cards, of which the four cards shown are just a sample; this is the intuition that lies behind Oaksford

and Chater's Bayesian approach. We will return to this interpretation in section 4.3.1 below. The other extreme is that each card defines a domain of its own, i.e. the rule is to be evaluated with respect to each card independently. The latter interpretation is the one suited to deontic conditionals, but there are indications that subjects sometimes impose this interpretation also in the indicative case, and then struggle with the resulting clash between two notions of truth. If a card complies with the rule, in other words 'if the rule is true of the card', then some subjects seem to have a tendency to transfer this notion of truth to 'truth of the rule *tout court*'. Here is an example of the phenomenon, observed in the two-rule task.

*Subject 10.*

*E.* If you found an 8 on this card [I], what would it say?

*S.* It would say that rule two is true, and if the two cannot be true then rule one is wrong...(Subject turns 8.)

*E.* OK so it's got an I on the back, what does that mean?

*S.* It means that rule two is true.

*E.* Are you sure?

*S.* I'm just thinking whether they are exclusive, yes because if there is an I then there is an 8. Yes, yes, it must be that.

One experimental manipulation in the tutorial dialogue for the two-rule task addressed this problem by making subjects first turn U and I, to find 8 on the back of both. This caused great confusion, because the subjects' logic (transferring 'truth of the card' to 'truth of the rule') led them to conclude that therefore both rules must be true, contradicting the instruction.

*Subject 18* [Initial choice was 8.]

*E.* Start with the U, turn that over.

*S.* U goes with 8.

*E.* OK now turn the I over.

*S.* Oh God, I shouldn't have taken that card, the first ...

*E.* You turned it over and there was an 8.

*S.* There was an 8 on the other side, U and 8. If there is an I there is an 8, so they are both true. [Makes a gesture that the whole thing should be dismissed.]

*Subject 28.*

*E.* OK turn them.

*S.* [turns U, finds 8] So rule one is true.

*E.* OK for completeness' sake let's turn the other cards as well.

*S.* OK so in this instance if I had turned that one [I] first then rule two would be true and rule one would be disproven. Either of these is different. [U or I]

*E.* What does that actually mean, because we said that only one of the rules could be true. Exactly one is true.

*S.* These cards are not consistent with these statements here.

On the other hand subjects who ultimately got the two-rule task right also appeared to have an insight into the intended relation between rule and cards.

*Subject 6.*

*E.* So say there were a U on the back of the 8, then what would this tell you?

*S.* I'm not sure where the 8 comes in because I don't know if that would make the U-one right, because it is the opposite way around. If I turned that one [pointing to the U] just to see if there was an 8, if there was an 8 it doesn't mean that rule two is not true.

We claim that part of the difficulty of the standard task involving a descriptive rule is the possibility of confusing the two relations between rule and cards. Transferring the 'truth of the card' to the 'truth of the rule' may be related to what Wason called 'verification bias', but it seems to cut deeper. One way to transfer the perplexity unveiled in the above excerpts to the standard task is to do a tutorial experiment where the A has a 4 on the back, and 7 an A. If a subject suffering from a confusion about the relation between cards and rule turns the A and finds 4, he should conclude that the rule is true, only to be rudely disabused upon turning 7. Unfortunately we haven't yet done this manipulation. In any case it is clear that for a deontic rule no such confusion can arise, because the truth value of the rule is not an issue.

### Exceptions and brittleness

The intended concept of truth is that of 'true without exceptions', what we call a brittle interpretation of the conditional. It goes without saying that this is not how a conditional is interpreted in real life. And we do find subjects who struggle with the required transition from a notion of truth where the exception may 'prove the rule', to exceptionless truth.

*Subject 18.*

*E.* What could you say is on the back of the 3, are you sticking with the consonant?

*S.* Consonant or U.

*E.* OK.

*S.* [Turns 3 and finds U] OK.. well no...well that could be an exception

you see.

*E.* The U?

*S.* The U could be an exception to the other rule.

*E.* To the first rule?

*S.* Yes, it could be an exception.

*E.* So could you say anything about the rule based on this? Say, on just having turned the U and found a 3?

*S.* Well yes, it could be a little exception, but it does disprove the rule so you'd have to... *E.* You'd have to look at the other ones? *S.* Yes.

Similarly in the standard Wason task:

*Subject 18.*

*S.* If I just looked at that one on its own [7/A] I would say that it didn't fit the rule, and that I'd have to turn that one [A] over, and if that was different [i.e. if there wasn't an even number] then I would say the rule didn't hold.

*E.* So say you looked at the 7 and you turned it over and you found an A, 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 A.

Clearly, if a counterexample is not sufficient evidence that the rule is false, then it is dubious whether card-turnings can prove the rule to be true or false at all. Subjects may accordingly be confused about how to interpret the instructions of the experiment. In any case a  $\neg q$  card would lose some salience (if it had any to begin with).

### The cards as sample

Above we noted that there are problems concerning the domain of interpretation of the conditional rule. The intended interpretation is that the rule applies to the four cards shown only. However, the semantics of conditionals is such that they tend to apply to an open-ended domain of cases. This can best be seen in contrasting universal quantification with the natural language conditional. Universal quantification is equally naturally used in framing contingent contextually determined statements as open-ended generalisations. So, to develop Goodman's (19??) example, "All the coins in my pocket this morning are copper" is a natural way to phrase a local generalisation with a fixed domain of interpretation. However, "If a coin is in my pocket this morning, it's copper" is a distinctly unnatural way of phrasing the same claim. The latter even invites the fantastical interpretation that if

a silver coin were put in my pocket this morning it would become copper—that is an interpretation in which a larger open-ended domain of objects is in play.

Similarly in the case of the four card task, the clause that “the rule applies only to the four cards” has to be explicitly included. One may question whether subjects take this clause on board, since this interpretation is an unnatural one for the conditional. A much more natural one is that the four cards are a sample. Indeed this is the point of purchase of Oaksford & Chater’s proposals that performance is driven by subjects’ assumptions about the larger domain of interpretation. We do find subjects who think that truth or falsity can only be established by (crude) probabilistic considerations:

*Subject 26.*

*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.

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.

Summarising: natural language descriptive conditionals bear complex relations to cases and sets of cases in their domain. In principle, only *sets* of cases can make a descriptive rule true. Even then the fact that all cases comply may intuitively not be enough, for instance when a subject hesitates to conclude ‘true’ from ‘not false’. The situation is still more complex because descriptive rules usually tolerate some exceptions. To get Wason’s desired interpretation of the rule as a material conditional, it is necessary to background the complex range of possibilities for descriptive rules’ relations to compliant cases and to exceptions, and to induce the intended meaning of ‘true’ and ‘false’. Here the two-rule task may have a role to play. If

subjects were assured that one of two rules was false and one was true, and instructed that their task was to gather minimal evidence as to which rule was which, then this hopefully focusses their attention on the more straightforward relations between rules and cases, and backgrounds the higher-order issues about how exceptions affect the truth of rules, and more generally the nature of truth. Of course the excerpts given above have mainly illustrated subjects' difficulties in the two-rule task. However, several tutorial dialogues involving the two-rule task also showed (very gradual and faltering) progress toward insight, while this progress was absent in the dialogues involving the standard task. This gave us some confidence that the two-rule task might be helpful in reaching the competence response, a prediction borne out by the experimental results reported below.

### 4.3.2 Dependencies between card-choices

The tutorial dialogues suggest that part of the difficulty of the selection task consists in having to choose a card *without being able to inspect what is on the other side of the card*. This difficulty can only be made visible in the dialogues because there the subject is confronted with real cards, which she is not allowed to turn at first. It then becomes apparent that some subjects would prefer to solve the problem by 'reactive planning', i.e. by first choosing a card, turning it and deciding what to do based on what is on the other side. This source of difficulty is obscured by the standard format of the experiment. The form invites the subjects to think of the cards depicted as real cards, but at the same time the answer should be given on the basis of the representation of the cards on the form, i.e. with inherently unknowable backs. The instruction 'Tick the cards you want to turn ...' clearly does not allow the subject to return a reactive plan. This is a pity, because the tutorials amply show that dependencies are a source of difficulty. Here is an excerpt from a tutorial dialogue in the two-rule condition.

*Subject 1.*

*E.* Same for the I, what if there is an 8 on the back?

*S.* If there is an 8 on the back, then it means that rule two is right and rule one is wrong.

*E.* So do we turn over the I or not?

*S.* Yes. Unless I've turned the U already.

And in a standard Wason task:

*Subject 10.*

*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.

Once one has understood Wason's intention in specifying the task, it is easy to assume that it is obvious that the experimenter intends subjects to

decide what cards to turn *before* any information is gained from any turnings. Alternatively, and equivalently, the instructions can be interpreted to be to assume the minimal possible information gain from turnings. However, the obviousness of these interpretations is possibly greater in hindsight, and so we set out to test whether they are a source of difficulty in the task. Note that no contingencies of choice can arise if the relation between rule and cards is interpreted deontically. Whether one case obeys the law is unconnected to whether any other case does. Hence the planning problem indicated above cannot arise for a deontic rule, which might be one explanation for the good performance in that case.

In this connection it may be of interest to consider the so-called *reduced array selection task*, or RAST for short, due to Wason and Green. In its barest outline<sup>6</sup> the idea of the RAST is to remove the  $p$  and  $\neg p$  cards from the array of cards shown to the subject, thus leaving only  $q$  and  $\neg q$ . The  $p$  and  $\neg p$  cards cause no trouble in the standard task in the sense that  $p$  is chosen almost always, and  $\neg p$  almost never, so one would expect that their deletion would cause little change in the response frequencies for the remaining cards. Surprisingly however, the frequency of the  $\neg q$  response increases dramatically. From our point of view, this result is perhaps less surprising, because without the possibility to choose  $p$ , dependencies between card choices can no longer arise.

### Getting evidence for the rule versus evaluation of the cards

A related planning problem, which can however occur only on a non-standard logical understanding of the problem, is the following. In a few tutorial dialogues involving the two-rule experiment, the background rule incorrectly failed to specify that the cards have one side either U or I and on the other side either 3 or 8. In this case the competence response is not to turn 3 only, but to turn U, I and 3. But several subjects did not want to choose the 3 for the following reason.

*Subject 7.*

*S.* Then I was wondering whether to choose the numbers. Well, I don't think so because there might be other letters [than U,I] on the other side. There could be totally different letters.

*E.* You can't be sure?

*S.* I can't be sure. I can only be sure if there is a U or an I on the other side. So this is not very efficient and this [3] does not give me

---

<sup>6</sup>The actual experimental set up is much more complicated and not quite comparable to the experiments reported here.

any information. But I could turn the U or the I.

Apparently the subject thinks that he can choose between various sets of cards and the choice should be as parsimonious as possible in the sense that every outcome of a turning must be relevant. To show that this is not an isolated phenomenon, here is a subject engaged in a standard Wason task:

*Subject 5.*

*E* So you would pick the A and you would pick the 4. And lastly the 7?

*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 A or a K again [writing the options for the back of 7 down], so if the 7 would have an A then that would prove me wrong. But if it would have a K then that wouldn't tell me anything.

*E.* So?

*S.* So these two [pointing to A and 42] give me more information, I think.

*E.* [...] You can turn over those two [A and 4].

*S.* [turns over the A]

*E.* So what does that say?

*S.* That it's wrong.

*E.* And that one [4]?

*S.* That it's wrong.

*E.* Now turn over those two [K and 7].

*S.* [Turning over the K] It's a K and 4. Doesn't say anything about this [pointing to the rule]. [After turning over the 7] 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 A, it could give me an a K and that wouldn't tell me anything.

*E.* But even though it could potentially give you an A on the back of it like this one has.

*S.* Yes, but that's just luck. I would have more chance with these two [referring to the A and the 4].

These subjects have no difficulty evaluating the meaning of the possible outcomes of turning 3 (in the two-rule task), or 7 (in the standard Wason task), but their choice is also informed by other considerations, in particular a perceived trade-off between the 'information value' of a card and the penalty incurred by choosing it. Of course this does not yet explain the evaluation of the 4/K card as showing that the rule is wrong, and simultaneously taking the K/4 card to be irrelevant. The combined evaluations

seem to rule out a straightforward biconditional interpretation of the conditional, and also the explanation of the choice of 4 as motivated by a search for confirmatory evidence for the rule, as Wason (and Oaksford & Chater) would have it. This pattern of evaluations is not an isolated phenomenon, so an explanation would be most welcome. Even without such an explanation it is clear that the problem indicated, how to maximise information gain from turnings, cannot play a role in the case of deontic conditionals, since the status of the rule is not an issue.

### 4.3.3 The pragmatics of the descriptive selection task.

The descriptive task demands that subjects seek evidence for the falsity of a statement which comes from the experimenter. The experimenter can safely be assumed to know what is (or is deemed to be) on the back of the cards. If the rule is false its appearance on the task sheet is the utterance, by the experimenter, of a knowing falsehood, possibly with intention to deceive. It is an active possibility that doubting the experimenter's veracity is a socially uncomfortable thing to do.

Quite apart from possible social psychological effects of discomfort, the communication situation in this task is bizarre. The subject is first given one rule to the effect that the cards have letters on one side and numbers on the other. This rule they are supposed to take on trust. Then they are given another rule by the same information source and they are supposed *not* to trust it but seek evidence for its falsity. If they do not continue to trust the first rule, then their card selections should diverge from Wason's expectations. If they simply forget about the background rule, the proper card choice would be A,K and 7; and if they want to test the background rule as well as the foreground rule, they would have to turn *all* cards. Notice that with the deontic interpretation, this split communication situation does not arise. The law stands and the task is to decide whether some people other than the source obey it. Here is an example of a subject who takes both rules on trust:

*Subject 3.* [Standard Wason task; has chosen A and 4] *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.

Another subject was rather bewildered when upon turning A he found a 7:

*Subject 8.*

*S.* Well there is something in the syntax with which I am not clear because it does not say that there is an exclusion of one thing, it says 'if there is an A on one side there is a 4 on the other side'. So the rule is wrong.

*E.* This [pointing to A] shows that the rule is wrong.

*S.* Oh so the rule is wrong, it's not something I am missing.

Although this may sound similar to Wason's 'verification bias', it is actually very different. Wason assumed that subjects would be in genuine doubt about the truth value of the rule, but would then proceed in an 'irrational', verificationist manner to resolve the issue. What transpires here is that subjects take it on the authority of the experimenter that the rule is true, and then interprets the instructions as indicating those cards which are evidence of this:

*Subject 22.*

*S.* Well my immediate [inaudible] first time was to assume that this is a true statement, therefore you only want to turn over the card that you think will satisfy the statement.

The communicative situation of the two-rule task is already much less bizarre, since there is no longer an reason to doubt the veracity of the experimenter. The excerpts also suggest that a modified standard task in which the rule is attributed not to the experimenter but to an unreliable source, might increase the number of competence responses. It hardly needs emphasising anymore that these problems cannot arise in the case of a deontic rule.

#### 4.3.4 Subjects' understanding of propositional connectives

As mentioned before, the tutorial dialogues were preceded by a paraphrase task, in which subjects were asked whether a statement involving a conditional is equivalent to a statement involving other logical connectives. A further striking observations from the paraphrase task is that a conditional  $p \rightarrow q$  is often (*i.* 50%) interpreted as a conjunction  $p \wedge q$ . Here is an example of what a conjunctive reading means in practice.

*Subject 22.* [Subject has chosen the conjunctive reading in the paraphrase task.]

*E.* [Asks subject to turn the 7]

*S.* That one ... that isn't true. There isn't an A on the front and a 4 on the back. [...] you turn over those two [A and 4] to see if they satisfy it, because you already know that those two [K and 7] don't satisfy the statement.

*E.* [baffled] Sorry, which two don't satisfy the rule?

*S.* These two don't [K and 7], because on one side there is K and that should have been A, and that [7] wouldn't have a 42, 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.

Clearly, on a conjunctive reading, the rule is already falsified by the cards as exhibited; no turning is necessary. The subject might however feel forced by the experimental situation to select some cards, and accordingly reinterprets the task as *checking* whether a given card satisfies the rule. This brings us to an important consideration: how much of the problem is caused by the conditional?

The literature on the selection task, with very few exceptions, has assumed that the problem is a problem specific to conditional rules. Indeed, it would be easy to infer also from the foregoing discussion of descriptive conditional semantics that the conditional (and its various expressions) is unique in causing subjects so much difficulty in the selection task, and that our only point is that a sufficiently rich range of interpretations for the conditional must be used to frame psychological theories of the selection task.

However, the issues already discussed—the nature of truth, response to exceptions, contingency, pragmatics—are all rather general in their implications for the task of seeking evidence for truth. One can distinguish the assessment of truth of a sentence from truthfulness of an utterer for sentences of any form. The robustness or brittleness of statements to counterexamples is an issue which arises for any generalisation. The social psychological effects of the experimenter's authority, and the communicative complexities introduced by having to take a cooperative stance toward some utterances and an adversarial one toward others is also a general problem of pragmatics that can affect statements of any logical form. Contingencies between feedback from early evidence on choice of subsequent optimal evidence seeking are general to any form of sentence for which more than one case is relevant.

It would seem to be a high priority to find out to what extent there is something uniquely problematic about conditionals in the selection task, and to what extent these more general issues could explain poor performance in seeking evidence for descriptive statements' truth. Several early papers compared disjunction with the conditional (e.g. van Duyne [107]), but disjunction has its own complexities and is closely allied to the conditional. What would happen, for example, if the rule were stated using the least problematical connective, conjunction?

## 4.4 Experiment

In this experiment, a number of conditions are compared with base-line performance on the classical descriptive ‘abstract’ task. We describe each condition in turn, and then present the results together.

### 4.4.1 The Conditions

#### Classical ‘abstract’ task

To provide a baseline of performance on the selection task with descriptive conditionals, the first condition repeats Wason’s (1968) classical study with the following instructions and materials:

Below is depicted a set of four cards, of which you can see only the exposed face but not the hidden back. On each card, there is a number on one of its sides and a letter on the other.

Also below there is a rule which applies only to the four cards. Your task is to decide which if any of these four cards you *must* turn in order to decide if the rule is true. Don’t turn unnecessary cards. Tick the cards you want to turn.

**Rule:** *If there is a vowel on one side, then there is an even number on the other side.*

#### Cards:



The other conditions are described through their departures from this condition.

#### Two-rule task

After the preliminary instructions for the classical task, the following instructions were substituted in this condition:

... Also below there appear two rules. One rule is true of all the cards, the other isn’t. Your task is to decide which cards (if any) you *must* turn in order to decide which rule holds. Don’t turn unnecessary cards. Tick the cards you want to turn.

**Rule 1:** *If there is a vowel on one side, then there is an even number on the other side.*

**Rule 2:** *If there is a consonant on one side, then there is an even number on the other side.*

Normative performance in this task, according to the classical logical competence model, is to turn only the not-Q card. The rules are chosen so that the correct response is to turn exactly the card that the vast majority of subjects fail to turn in the classical task. This has the added bonus that it is no longer correct to turn the P card which provides an interesting comparison with the classical task. This is the only descriptive task for which choosing the true-antecedent case is an error. Mental models theory supposes that subjects always tend to reason about this case first.

By any obvious measure of task complexity, this task is more complicated than the classical task. It demands that two conditionals are processed and that the implications of each case is considered with respect to both rules and with respect to a distribution of truth values. The normative response is to turn *neither* the true-antecedent nor true-consequent cards. Nevertheless, our prediction was that performance should be substantially nearer the logically normative model, as long as this manipulation succeeded in backgrounding concerns about the significance of exceptions and some of the other semantic issues discussed above.

One might argue that a really persistent worrier about exceptions could still have problems in this task, since they could be confused about what would they should do if exceptions were found to *both* rules. As with all the experimental manipulations explored in this paper, there is no guarantee that instructional or task changes will have the desired clarificatory effects. Nevertheless, to the extent they do, the experiments provide evidence that a problem was significant in the classical task. We return in the discussion to the issue of interpreting the effects of this task change.

### Contingency instructions

The ‘contingency instructions’, designed to remove any difficulties from interpreting our intentions in this regard, after an identical preamble, read as follows, where the newly italicised portion is the change from the classical instructions:

... Also below there appears a rule. Your task is to decide which of these four cards you *must* turn (if any) in order to decide if

the rule is true. *Assume that you have to decide whether to turn each card before you get any information from any of the turns you choose to make. Don't turn unnecessary cards. Tick the cards you want to turn.*

If the contingencies introduced by the descriptive semantics are a source of difficulty for our subjects, this additional instruction should make the task easier. In particular, since there is a tendency to choose the P card first, there should be an increase in not-Q responding.

### Judging truthfulness of an independent source

We chose to investigate the possible contribution of problems arising from the authoritative position of the experimenter and the balance of cooperative and adversarial stances required toward different parts of the task materials through instructions to assess truthfulness of the source instead of truth of the rule, and we separated the source of the rule from the source of the instructions (the experimenter). The instructions read as follows:

... Also below there appears a rule *put forward by an unreliable source*. Your task is to decide which cards (if any) you *must* turn in order to decide *if the unreliable source is lying*. Don't turn unnecessary cards. Tick the cards you want to turn.

These instructions mean that the source of the rule is not the experimenter, and so there can be no discomfort about seeking to falsify the experimenter's claim. Nor should any falsity of the rule throw any doubt on the truthfulness of the rest of the instructions, since the information sources are independent.

These 'truthfulness' instructions are quite closely related to several other manipulations that have been tried in past experiments. In the early days of experimentation on this task, when it was assumed that a failure to try and falsify explained the correct response, Griggs [38], concerned with Wason's ideas about subjects seeking positive cases, explored instructions to falsify the rule (rather than find out whether it was true). This manipulation produced no increase in not-Q responding. But note that this instruction removed neither of the problems we focus on here.

Wason himself compared instructions to test 'whether the rule is lying' and found that this made little difference. However, this instruction fails to separate the source of the rule from the experimenter (as the utterer of the rule) and may fail for that reason. Kirby used a related manipulation

in which the utterer of the rule was a machine said to have broken down, needing to be tested to see if it was working properly again after repair. These instructions did produce significant improvement. Here the source is unreliable but one might argue that the focus of the instruction is to tell whether the machine is ‘broken’, not simply whether the utterance of the rule is a falsehood. This might be expected to invoke a deontic interpretation (Kirby’s condition is akin to the ‘production line inspection scenarios’ mentioned before), and so it might be that the improvement observed is for this reason.

It appears that the present manipulation has not been explored before. We predicted that separating the source of the rule from the experimenter while maintaining a descriptive reading of the rule should increase normative responding.

### Exploring other kinds of rules than conditionals

This condition of the experiment was designed to explore the malleability of subjects’ interpretations of rules other than conditionals. In particular we chose a conjunctive rule as arguably the simplest connective to understand. As such this condition has a rather different status from the others in that it is not designed to remove a difficulty from a logically similar task but to explore a logical change. Since it was an exploration we additionally asked for subjects’ justification of their choices afterwards.

A conjunctive rule was combined with the same instructions as are used in the classical abstract task.

**Rule:** *There is a vowel on one side, and there is an even number on the other side.*

The classical logical competence model demands that subjects should turn no cards with such a conjunctive rule—the rule interpreted in the same logic as Wason’s interpretation of his conditional rule can already be seen to be false of the not-P and not-Q cards. Therefore, under this interpretation the rule is already known to be false and no cards should be turned.

We predicted that many subjects would not make this interpretation of this response. An alternative, perfectly rational, interpretation of the experimenter’s intentions is to construe the rule as having deontic force (every card *should* have a vowel on one side and an even number on the other) and to seek cards which might flout this rule other than ones that

obviously can already be seen to flout it. If this interpretation were adopted, then the P and Q cards would be chosen. Note that this interpretation is deontic even though the rule is syntactically indicative.

#### 4.4.2 Subjects

Subjects were 377 first year Edinburgh undergraduates, from a wide range of subject backgrounds.

#### 4.4.3 Method

Subjects were randomly assigned to the different conditions, with the size of sample in each condition being estimated from piloting on effect sizes. All tasks were administered to subjects in a classroom setting. Students worked alone and were separated from other subjects doing the same task.

#### 4.4.4 Results

Those subjects (12 across all conditions) who claimed to have done similar tasks before, or to have received any instruction in logic were excluded from the analysis.

Table 4.1 presents the data from all of the conditions. Any response made by at least three subjects in at least one condition is categorised: all other responses are treated as miscellaneous.

Condition	P Q	Q	P	P $\neg$ Q	$\neg$ Q	$\neg$ P,Q	P,Q, $\neg$ Q	$\neg$ P, $\neg$ Q	all	None	Misc.	Tot
Classical	56	7	8	4*	3	7	1	2	9	8	5	108
Truthfulness	39	6	9	14*	0	7	3	6	8	15	5	112
2-rule	8	8	2	1	9*	2	1	0	0	2	4	37
Contingency	15	0	3	8*	1	6	4	8	3	0	3	51
Conjunctive	31	2	9	7	2	0	0	1	0	9*	8	69

Table 4.1: Frequencies of card choice combinations by conditions. Classical logical competence responses are marked \*. Any response made by at least three subjects in at least one condition is categorised: everything else is miscellaneous

Subjects were scored as making a completely correct response, or as making at least some mistake, according to the classical logical competence

model. For all the conditions except the two-rule task and the conjunction condition, this ‘competence model’ performance is choice of P and not-Q cards. For the two-rule task the correct response is Q. For the conjunction condition it is to turn no cards.

Table 4.1 presents the tests of significance of the percentages of correct/incorrect responses as compared to the baseline classical condition. 3.7% of subjects in the baseline condition made the correct choice of cards. This is in line with reported results for this task, toward the lower end. For example, Hoch & Tschirgi (1985) report 4% correct for a comparable sample of highschool students (Scottish first-year undergraduates are a year younger than UK and US students, and these were tested at the beginning of their first year).

The percentages completely correct in the other conditions were 2-rule condition 24%; ‘truthfulness’ condition 13%; in the ‘contingency’ condition 18%; and in the conjunction condition 13%. The significance levels of these proportions by Fisher’s exact test appear in Table 4.2.

Condition	Wrong	Right	p	Percent Correct
Classical baseline	104	4		3.7
2-Rule	28	9	.004	24
Truthfulness	98	14	.033	13
Contingency	37	8	.005	18
Conjunction	60	9	.022	13

Table 4.2: Proportions of subjects completely correct and significances of differences from baseline of each of the four manipulations.

For subjects, the two-rule task is substantially easier than the baseline task. In fact the completely correct response is the modal response. More than six times as many subjects get it completely correct even though superficially it appears a more complicated task. The next most common responses are to turn P with Q, and to turn just Q. The former is the modal response in the classical task. The latter appears to show that even with unsuccessful subjects, this task shifts attention to the consequent cards—turnings of P are substantially suppressed 32% as compared to 80% in the baseline task.

Contingency instructions also substantially increase completely correct responding, and do so primarily at the expense of the modal P with Q response. In particular they increase not-Q choice to 50%.

Instructions to test the truthfulness of an unreliable source have a smaller effect which takes a larger sample to demonstrate, but nevertheless, 13% of subjects get it completely correct, nearly four times as many as the baseline task. The main change is again a reduction of P with Q responses, but there is also an increase in the response of turning nothing.

Completely correct performance with a conjunctive rule was 13%—not very different from the conditions with conditional rules. The modal response is to turn the P and Q cards—just as in the original task. Anecdotally, debriefing subjects after the experiment suggests that at least a substantial number of modal responses are explained by the subjects in terms construable as a deontic interpretation of the rule, roughly paraphrased as “The cards should have a vowel on one side and an even number on the other”.

## 4.5 The meaning of conditionals

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* [18], 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’)<sup>7</sup>, 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 [17], the authors observed that there occurred two main types of conditionals,

---

<sup>7</sup>Dutch, however, can also use the conditionals marker ‘als’ here.

*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. ([17],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 [18])

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 ([17],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 5.1 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 paraphrase 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 [27] already observed that there are high frequencies 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.

## 4.6 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. Similarly, adopting this interpretation together with a biconditional reading could explain the selection of the  $p, q$  cards. Johnson-Laird and Wason [58] referred to this phenomenon by saying that subjects do not always recognise the *reversibility* of the cards. In another paper, Wason and Johnson-Laird

[113] 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 [31] have used a modified method to argue that constant 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.<sup>8</sup>

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 sub-

---

<sup>8</sup>Four combinations, because the constant back/face reading of the anaphor appears to be too implausible to be considered.

jects 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.

The data are presented in Appendix A, section 4.6.1. 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 4.6. 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. The conditions used in this experiment were as follows. We tried to induce correct understanding of the anaphoric expression ‘one side - - other side’ by treating the three possibilities explicitly

1. if there is a vowel on the (visible) face, then there is an even number on the (invisible) back
2. if there is a vowel on the (invisible) back, then there is an even number on the (visible) face
3. 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. These conditions will be referred to as experiments 1, 2, 3 respectively.

We now provide a number of examples, culled from the tutorial dialogues, 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 1,2,3]

*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 3, but nonetheless goes astray when forced to consider the back/face interpretation.

*Example* Subject 4.[experiments 1,2,3]

*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 1,2,3]

[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* [Standard Wason task]

*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*.

[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 and pragmatics, 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 combined semantic and pragmatic explanation could refer to the view that conditionals with consequents known to be true are odd, some would even say ungrammatical (Haiman [41]). 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<sup>9</sup> Because these conditionals are pragmatically perverse, they may be subject to what Fillenbaum [27] 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.

---

<sup>9</sup>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.

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

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*<sup>10</sup>.

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, 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 [87] and Gebauer and Laming [31]. Their ‘static’ analysis may of course apply to some subjects, but the excerpts presented above seem to require a more dynamic analysis.

With hindsight, one can see that the issue of anaphora was implicitly raised by Wason and Green [112], although their focus is on the distinction between a *unified* and a *disjoint* representation of the stimulus. A unified

---

<sup>10</sup>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.

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.<sup>11</sup>

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 [112], p. 604–607).

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 2 with its constant back/face anaphora, where the object mentioned in the antecedent is hidden. We have seen in section 4.6 that some subjects have difficulties with the intended direction of the conditional in experiment 2.

---

<sup>11</sup>A RAST sometimes improves performance, e.g. in case of sentence (2a) below. However, our condition 2 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.

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 3; 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’ [112, 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

Response	Instruction				Total
	classical	reversible	constant-front	constant-back	
0000	3	2	3	0	8
0001	0	0	0	0	0
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 4.3: Frequencies of responses for the four different rule in the conditions 4.6. ‘0’ in the response label indicates no turn; ‘1’ indicates subject turns card (cards in the usual  $p, \neg p, q, \neg q$  order).

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 [96] for an approach to this question); Wason and Green confess they have little to offer here.

#### 4.6.1 Appendix A: Classroom Experiment Data

Table 4.3 displays the data from the initial classroom experiment comparing selections for four different rules as specified in 4.6.

## Chapter 5

# Logic and probability

Probabilistic approaches to the selection task aim to show that ‘irrational’ answer usually given is actually rational according to a probabilistic (Bayesian) competence model. However, it is a matter of dispute whether humans do reason in conformity with Bayesian norms.

### 5.1 Probability underlies logic

#### Bayesian explanations of reasoning

We will now discuss in somewhat greater detail the Bayesian explanation of behaviour in the 4-card task due to Oaksford and Chater [77], 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.<sup>1</sup>

---

<sup>1</sup>This reading could also be suggested by an analysis of the conditional along the lines proposed by Lewis [68] which in our case would run as follows: ‘(Always: if x has a vowel



*E.* [...] You can turn over those two [A and 4].  
*S.* [turns over the A]  
*E.* So what does that say?  
*S.* That it's wrong.  
*E.* And that one [4]?  
*S.* That it's wrong.  
*E.* Now turn over those two [K and 7].  
*S.* It's a K and 4. Doesn't say anything about this [pointing to the rule]. [After turning over the 7] 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 A, it could give me an a K and that wouldn't tell me anything.  
*E.* But even though it could potentially give you an A 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 A and the 4].

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 [77].<sup>2</sup> 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 utilities further.

---

<sup>2</sup>Not unequivocally, however, because, as we have seen the  $q$  card may be selected for its potentially falsifying, not verifying, character.

We will now discuss the model in greater formal detail. Interestingly, it is adapted (cf. Oaksford and Chater [78]) 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  $p \rightarrow q$ . It is fundamental to Oaksford and Chater’s [77] reconstruction that they assume that a subject interprets the conditional as pertaining to a population from which the four cards shown are only a sample. Of course, this was not the way the task was 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  $P(p)$ ,  $P(q)$ <sup>3</sup>, as follows:

1. if  $P(p)$ ,  $P(q)$  are small ( $\leq 0.15$ ),  $E_p(I) > E_q(I) > E_{\neg q}(I) > E_{\neg p}(I)$ ;
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

---

<sup>3</sup>Strictly speaking one also has dependence on  $P(H_0)$  but the rank order is by and large independent of this value.

the  $q$  card over turning the  $\neg q$  card.<sup>4</sup>

We performed the experiment involving two rules, one true, one false, 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  $p \rightarrow \neg q$ .

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 [66] 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, i.e. the assumption that both  $P(p)$  and  $P(q)$  are small. Take the case

---

<sup>4</sup>It is somewhat peculiar that Oaksford and Chater [78] refer to Horwich’ *Probability and evidence* [51] 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.

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 [77] offer two solutions here. The first derives from Oaksford and Stenning [76] 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  $\leq 0.5$ . 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  $P(q) \geq P(\sim p)P(H'_1)$ , 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 [78] reaction to an experiment of Pollard and Evans (for a discussion, see Evans

and Over [25]), which at least at first sight appears to be a test of this particular Bayesian model. Pollard and Evans manipulated the conditional probability  $P(q|p)$  (which they equate with the probability of the conditional  $p \rightarrow q$ ) with a view to demonstrating that if the conditional is usually false, i.e. if  $P(q|p)$  is low, then subjects are more likely to choose the  $p, \neg q$  cards. The manipulation consisted in showing subjects two sets of cards. One set (for the 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 loses 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 [107]. 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<sup>5</sup>,

---

<sup>5</sup>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  $p \rightarrow q$  is not perceived as being equivalent to  $\neg p \vee q$  and furthermore that it is hard to process the latter form.

in the sense that the percentage of correct answers is the same.<sup>6</sup> 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.

---

<sup>6</sup>Interestingly, the 'matching response'  $p, q$  occurs much less frequently in formulation 3.



## Chapter 6

# Logic and evolution

As we have seen there have been recurrent attempts to oppose form and content in logical reasoning, usually to the detriment of the former. The most forceful assault on form is due to the evolutionary psychologist Leda Cosmides, who claims that (1) successful reasoning (according to the canons of classical logic) occurs only in the case of narrowly circumscribed contents, and (2) these contents have been selected by evolution, i.e. they correspond to situations in our environment to which we are especially attuned because they are crucial for survival.

Originally, Cosmides thought that all successful reasoning involves *social contracts*, in which parties agree to exchange benefits. Recently, the catalogue has been extended to include reasoning about precautions, warnings etc., but our discussion will concentrate on the original idea; nothing is lost thereby.

### 6.1 Evolutionary thinking: from biology to evolutionary psychology.

We assume the reader is familiar with the broad outlines of the theory of evolution. To state precisely what the theory of evolution is, immediately involves us in a number of conceptual puzzles. For instance: what is selected? Genes, cells, traits, organisms (phenotypes), groups of organisms?

Evolutionary psychologists believe that human cognition is *in toto* a product of evolution. This claim may seem counterintuitive at first sight, since humans have impressive learning abilities.

Human cognition consists of a number of adaptations, which have been useful for survival at one stage or another of the Environment of Evolution-

ary Adaptation (EEA), the Pleistocene, when (proto-) humans lived in small bands as hunter-gatherers. Since this time period is very much longer than the agricultural period (which started some 12.000 years ago), such traits as humans have must either have evolved as part of their common ancestry with primates, or in the EEA.

In this respect the hotly debated issue of the origin of language is of particular interest. As is well known, Chomsky advanced the *poverty of the stimulus* argument to substantiate his view that the language capacity must somehow be innate. In outline, the argument is this: children reach grammatical competence in language very quickly, at about age 8. This competence includes complicated constructions of which they will have heard few examples; more importantly, the recursive nature of language cannot be inferred only from the linguistic data presented to the child. But if that is so, some construction principles ('Universal Grammar') must be innate; the role of the linguistic input is to set some parameters in universal grammar (hence 'principles and parameters'). Chomsky wishes to remain agnostic about the possible evolutionary origins of universal grammar. Some followers (Pinker, more recently Jackendoff) have been less inhibited and have argued that humans must have evolved a specific 'language module', a genetically determined piece of wetware specifically dedicated to the processing of language. They argue that Chomsky should have followed his own argument to its logical conclusion, by identifying 'innate' with 'genetically determined'. This leads us into two interesting issues: (1) is the identification of 'innate' with 'genetically determined' really unproblematic?, and (2) what is this notion of a 'module'?

### 6.1.1 Adaptations and exaptations

A 'trait' is an aspect of an organism's phenotype, particularly an aspect that can have effects in the environment. This definition is not very operationalised, but is nevertheless fundamental to a discussion of evolution.

If a trait is genetically determined, then the trait evolves as the genes from which it develops evolve. Four factors may influence the evolution of genes: mutation, migration, drift, and (natural or sexual) selection. An *adaptation* is a trait that has been selected under pressure from natural selection for a particular gene-propagating effect. That effect is then called the function of the trait. A prime example is the eye.

Opposed to the adaptation is the *exaptation*: an evolved trait that acquires a new (beneficial) effect, without having been selected for that particular effect. Here the prime example is birds' feathers: 'designed' for

temperature regulation, they were later exapted for flight. (This is *primary exaptation*. A process of *secondary adaptation* may have led to further improvements in the flight-enabling properties of feathers.) A slightly different example is furnished by snails that use a space in their shell to brood eggs. The existence of that space is a necessary consequence of the spiral design of shells—but only the latter can be said to have evolved by selection. Apparently the snails that use the space for brooding eggs evolved after the ones who don't: in this sense the open space is exapted to brooding eggs. (Such exaptations are often called *spandrels*, an architectural term referring to the open surfaces on the ceiling of a Gothic church, necessarily created by its supporting arches. These surfaces are often decorated with mosaics; but the point is, as argued by Gould and Lewontin, that they were not designed for that purpose.)

We thus see that there are several senses in which a trait (such as the language capacity) can be innate. Evolutionary psychologists tend to hold that all traits of interest, including cognitive traits, are adaptations. In fact, they claim that only an evolutionary perspective can give cognitive science theories of any explanatory power, where 'evolution' is taken to be tantamount to 'shaped by natural selection'. Accordingly, they advocate that cognitive science's methodology should proceed as follows:

1. analyse the information-processing task (in the sense of Marr) that corresponds to a particular cognitive trait
2. find the highly specialised mechanism that is able to execute this task
3. explain how this mechanism came about as an adaptation

An organism's phenotypic structure can be thought of as a collection of 'design features'—micro-machines such as the functional components of the eye and the liver. The brain can process information because it contains complex neural circuits that are functionally organized. The only component of the evolutionary process that can build complex structures that are functionally organized is natural selection. And the only kind of problems that natural selection can build complexly organized structures for are adaptive problems, where 'adaptive' has a very precise, narrow technical meaning. Natural selection is a feedback process that 'chooses' among alternative designs on the basis of how well they function. By selecting designs on the basis of how well they solve adaptive problems, this process engineers a tight fit

between the function of a device and its structure. [C]ognitive scientists need to realize that while not everything in the designs of organisms is the product of selection, all complex functional organization is.

They thus have little time for possible structural features of cognition arising as spandrels. Their picture of the human mind is that of a collection of adaptations (‘a collection of ‘design features’’) more or less hanging together, *tant bien, tant mal*; in a very evocative metaphor: ‘the mind is a Swiss army knife’.

### 6.1.2 Massive modularity

This point of view has been aptly described as *massive modularity*: the mind is completely composed of domain-specific modules (‘Darwinian algorithms’)

[Content-specific mechanisms] will be far more efficient than general purpose mechanisms . . . [content-independent systems] could not evolve, could not manage their own reproduction, and would be grossly inefficient and easily out-competed if they did.

This form of modularity is to be contrasted with Fodorian modularity, which is more of a hybrid architecture. On the one hand there are low-level modules for input and output computations (primarily perception and motor control), but these are connected to a central processing unit, which is able to integrate information from the various low-level modules. As such, it must function in a modality-independent manner; perhaps by symbol-processing. That is, one picture of the functioning of the CPU is that the input is first translated into symbols, then processed, then translated back into motor commands. Evolutionary psychologists thus claim that Fodorian modularity would be less adaptive (would less lead to reproductive success) than massive modularity; they view the processing that would have to go on in the CPU as so much useless overhead.

### 6.1.3 No role for logic?

It follows from the above considerations that there is no room for domain-independent forms of learning, memory and reasoning. ‘Reasoning’ is simply the wrong abstraction: there exists ‘reasoning’ in specific domains, designed to solve a particular adaptive problem, but there is no general overarching innate capacity for reasoning. But by its very definition, logic seems to be

content-independent: an argument is valid if whatever is substituted for the nonlogical terms, true premises lead to a true conclusion. Hence logic must be an acquired trick: humans have no special capacity for formal reasoning. Indeed, the difficulty of mastering logic points to its lack of biological roots: the existence of an adaptive module is usually reflected in the ease with which humans learn to use it effectively and quickly.

#### 6.1.4 Cheater detection

The research on the Wason task and other reasoning tasks has shown that people reason correctly (according to the norms of classical logic). The argument outlined above suggests that there must be an adaptation, a specific module, responsible for this good performance. Cosmides proposed that logically correct reasoning has its origins in social cooperation. In particular, she focusses on the domain of social contracts, in which one party is willing to pay a cost to acquire a certain benefit from a second party. In order to ensure survival it would be imperative to be able to check for parties cheating on the contract, i.e. parties accepting the benefit without paying the associated cost.

For hunter-gatherers, social contracts, that is, cooperation between two or more people for mutual benefit, were necessary for survival. But cooperation (reciprocal altruism) cannot evolve in the first place unless one can detect cheaters (Trivers 1971). Consequently, a set of reasoning procedures that allow one to detect cheaters efficiently—a cheat-detector algorithm—would have been selected for. Such a ‘Darwinian algorithm’ would draw attention to any person who has accepted the benefit (did he pay the cost?) and to any person who has not paid the cost (did he accept the benefit?). Because these reasoning procedures, which were adapted to the hunter-gatherer mode of life, are still with us, they should affect present day reasoning performance.

The prediction of (Cosmides’ variant of) evolutionary psychology in a reasoning task is thus that performance will be only be good if cheater detection is activated.

#### Methodology

There is still a problem ‘prediction of S.M.S.S. in reasoning task: always applicable, hence always successful’

## 6.2 Logic and evolution—experimental data

### 6.2.1 Social contracts and cheating detection

Cosmides ran a number of experiments contrasting social contracts with arbitrary regulations, and contrasting cheater detection with altruism detection. Here is her famous experiment on cheater detection in a social contract.

You are an anthropologist studying the Kaluame, a Polynesian people who live in small, warring bands on Maku Island in the Pacific. You are interested in how Kaluame ‘big men’—chieftains—yield power.

‘Big Kiku’ is a Kaluame big man who is known for his ruthlessness. As a sign of loyalty, he makes his own ‘subjects’ put a tattoo on their face. Members of other Kaluame bands never have facial tattoos. Big Kiku has made so many enemies in other Kaluame bands, that being caught in another village with a facial tattoo is, quite literally, the kiss of death.

Four men from different bands stumble in Big Kiku’s village, starving and desperate. They have been kicked out of their respective villages for various misdeeds, and have come to Big Kiku because they need food badly. Big Kiku offers each of them the following deal: ‘If you get a tattoo on your face, then I’ll give you cassava root.’

Cassava root is a very sustaining food which Big Kiku’s people cultivate. The four men are very hungry, so they agree to Big Kiku’s deal. Big Kiku says the tattoos must be in place tonight, but that the cassava root will not be available until the following morning.

You learn that Big Kiku hates some of these men for betraying them to his enemies. You suspect he will cheat and betray some of them. Thus, this is the perfect opportunity for you to see first hand how Big Kiku wields his power.

The cards below have information about the fates of the four men. Each card represents one man. One side of a card tells whether or not the man went through with the facial tattoo that evening and the other side of the card tells whether or not Big Kiku gave that man cassava root the next day.

Did Big Kiku get away with cheating any of these four men?  
Indicate only those card(s) you definitely need to turn over to  
see if Big Kiku has broken his word to any of these four men.

got tattoo	got no tattoo	B.K. gave cassava	B.K. gave nothing
------------	---------------	-------------------	-------------------

75% of subjects now chose the *got tattoo* and *B.K. gave nothing* cards, a score comparable to that for the 'drinking age' rule. Cosmides interprets this result as falsifying an explanation of the alleged content effect based on familiarity with the rule. This explanation was advanced after the observation that whereas British students did well with the postal rule

If an envelope is sealed, it has a 3p stamp

students in the U.S., where such a rule is unknown, scored badly with this rule. This observation suggested to some that not only the concept of a rule, but the very content of the rule must be familiar. As opposed to this, Cosmides claims that also unfamiliar content may elicit good performance, as long as a social contract is involved.

It will be clear by now that our explanation of the result is very different: Cosmides' rule is of deontic nature, and hence none of the factors that complicated reasoning in the case of descriptive conditionals are operative here, and so one would expect many competence answers here.

**Why cheater detection matters more than logic** Cosmides claims that in some cases of reasoning about social contracts, the predictions of logic and cheater detection theory diverge; the experimental results then show that the former are falsified. Consider the following social contract:

- (1) If you give me your watch, I give you 20 euro.

According to Cosmides, this contract is equivalent to the following:

- (2) If I give you 20 euro, you give me your watch.

Indeed, both contracts express that the 'I' is willing to pay a cost (20 euro) in order to receive a benefit (the watch) from 'you'. The only difference appears to lie in the ordering of the transactions, in the sense that usually, but not always, the action described in the consequent follows the action described in the antecedent. Now suppose the following cards are laid out on the table

gave watch	didn't give watch	gave 20 euro	B.K. gave 10 euro
------------	-------------------	--------------	-------------------

then contracts (1) and (2) would, according to Cosmides, both lead to the choice ‘gave watch’ and ‘gave 10 euro’, since ‘I’ have cheated ‘you’ if ‘I’ accept ‘your’ watch while paying ‘you’ less than the 20 euro that we agreed upon.

Observe that if the contract is expressed in the form (1), the cards showing ‘gave watch’ and ‘gave 10 euro’ would correspond with the antecedent and the negation of the consequent of the conditional. If the contract is expressed in the form (2), these cards correspond instead with the consequent and the negation of the antecedent of the conditional. Cosmides now claims that the prediction of propositional logic is different from that of cheater detection, since a falsifying instance would always be of the form ‘antecedent plus negation of consequent’, whereas as we have seen instances of cheating can take different forms. Thus, if the contract is presented in the form (2), logic and cheater detection would dictate different answers.

Similarly, suppose that the deal proposed by Big Kiku is formulated as a *switched social contract*

If I give you cassava root, you must get a tattoo on your face.

Cosmides claims that (a) the original and the switched rule embody the same social contract, therefore in both cases the cards *got tattoo* and *B.K. gave nothing* would have to be chosen, and (b) logic dictates that in the case of the switched social contract the following cards would have to be chosen: *B.K. gave cassava* and *no tattoo*. The argument for this is that only these cards can yield counterexamples to the conditional as stated.

While we agree to (a), we consider (b) to be another example of the surface form fetishism that has so marred the subject. It is precisely because (a) is true that the logical form of either the original or the switched is not that of a material conditional.

**Varieties of deontic logic** The proper formalism in which to formulate social contracts is deontic logic, which is concerned with the logic of ‘ought’. There are two ways of conceiving ‘ought’: as a monadic operator  $Oq$ , or as a dyadic operator  $O(p, q)$ . In practice, obligations are used conditionally: if  $p$  is the case, then one should do  $q$ . In a monadic system, such obligations can be formulated as either  $p \rightarrow Oq$  or as  $O(p \rightarrow q)$ . The difference between these formulations can be seen once we introduce a semantics for  $O$ .

**Definition 4** A model structure for monadic deontic propositional logic is a triple  $M = (W, R, V)$  where  $W$  is a nonempty set of worlds,  $R \subseteq W \times W$

is the relation ‘having a deontically perfect alternative’, and  $V$  assigns a subset of  $W$  to each proposition letter.

If  $w$  is a world in  $M$ , we put  $w \Vdash O\psi$  iff for all  $v$  such that  $Rwv$ ,  $v \Vdash \psi$ . A number of conditions are put on  $R$  in order to ensure the right properties of  $O$ , but these are not our present concern.

Thus  $w \Vdash p \rightarrow Oq$  means: if  $p$  is true in  $w$ , then  $q$  ought to be the case in a deontically perfect alternative to  $w$ , and  $w \Vdash O(p \rightarrow q)$  means: if  $v$  is a deontically perfect alternative to  $w$  that satisfies  $p$ , then  $q$  ought to be the case there as well. The first formulation is often called an *absolute* commitment, the second a *prima facie* commitment. The difference is of course that in the latter case, from  $p$  and  $O(p \rightarrow q)$  it does not follow that  $Oq$ . For simplicity suppose therefore that we formulate conditional obligations as  $p \rightarrow Oq$ .

When there are several parties to a contract, one must introduce indexed deontic operators, one for each party, and corresponding indexed alternative relations. In case of a social contract between an ‘I’ and a ‘You’, the contract is formulated by means of the *two* formulas  $p \rightarrow O_I q$  and  $q \rightarrow O_Y p$ . This is just the formal translation of an observation due to Geis and Zwicky to the effect that a conditional promise often invites the inference to its converse (cf. [32]). In world  $w$ , a violation of the contract from the point of view of ‘You’ is a world  $v$  such that  $v \Vdash p, \neg q$  and  $\neg R_I wv$ , and a violation from the point of view of ‘Me’ is a world  $u$  such that  $u \Vdash \neg p, q$  and  $\neg R_Y wu$ .

Returning to Big Kiku’s village, we see that there exists a contract between Big Kiku and each of the men. Each such contract is of the form  $p \rightarrow O_{BK} q \wedge q \rightarrow O_m p$ , where  $p$  is ‘got tattoo’ and  $q$  is ‘Big Kiku gave cassava’. In both scenarios the perspective of one of the men is taken; but this means that in both cases attention is focussed on the first conditional and its deontically suboptimal worlds where  $p, \neg q$  is the case.

**Altruism** Cosmides’ second way of arguing that competence reasoning with ‘if... then’ is due only to the activation of cheater detection, and not to the *logical form* of social contracts, is to present an example of reasoning with social contracts in which humans don’t excel. Evolutionary theories would not require the existence of ‘altruists’, that is, individuals who are willing to pay a price without taking the corresponding benefit. These individuals would quickly lose out in the struggle for survival, and hence don’t exist. But if altruists don’t exist, there has been no need for an ‘altruist detector’ to evolve, and accordingly humans should not exhibit a special ability to reason about altruism with respect to a given social contract. The argument

outlined here already raises many questions, but let us take it for granted for the moment. We will show that the experiment designed to verify the prediction leaves much to be desired.

We proceed to give *in extenso* Cosmides' instructions, which are for the most part identical to those for the case of cheater detection.

You are an anthropologist studying the Kaluame, a Polynesian people who live in small, warring bands on Maku Island in the Pacific. You are interested in how Kaluame 'big men'—chieftains—yield power.

'Big Kiku' is a Kaluame big man who is known for his ruthlessness. As a sign of loyalty, he makes his own 'subjects' put a tattoo on their face. Members of other Kaluame bands never have facial tattoos. Big Kiku has made so many enemies in other Kaluame bands, that being caught in another village with a facial tattoo is, quite literally, the kiss of death. Four men from different bands stumble in Big Kiku's village, starving and desperate. They have been kicked out of their respective villages for various misdeeds, and have come to Big Kiku because they need food badly. Big Kiku offers each of them the following deal: 'If you get a tattoo on your face, then I'll give you cassava root.'

Cassava root is a very sustaining food which Big Kiku's people cultivate. The four men are very hungry, so they agree to Big Kiku's deal. Big Kiku says the tattoos must be in place tonight, but that the cassava root will not be available until the following morning.

You learn that Big Kiku hates some of these men for betraying them to his enemies. You suspect he will cheat and betray some of them. However, you have also heard that Big Kiku sometimes, quite unexpectedly, shows great generosity towards others - that he is sometimes quite altruistic. Thus, this is the perfect opportunity for you to see first hand how Big Kiku wields his power.

The cards below have information about the fates of the four men. Each card represents one man. One side of a card tells whether or not the man went through with the facial tattoo that evening and the other side of the card tells whether or not Big Kiku gave that man cassava root the next day.

Did Big Kiku behave altruistically towards any of these four men? Indicate only those card(s) you definitely need to turn

over to see if Big Kiku has behaved altruistically to any of these four men.

got tattoo	got no tattoo	B.K. gave cassava	B.K. gave nothing
------------	---------------	-------------------	-------------------

Cosmides claims that in this experiment the ‘no tattoo’ and ‘Big Kiku gave cassava’ cards would have to be turned, with the following argument.

Altruists, according to Cosmides, are people ready to pay a price without taking the corresponding benefit; if Big Kiku is an altruist, he wants to pay a price (give cassava root), without demanding his rightful benefit (the tattoo). Hence it would have to be checked whether behind the ‘no tattoo’ card is written ‘Big Kiku gave cassava’ and whether behind the ‘Big Kiku gave cassava’ card it is written kaart ‘no tattoo’. Very few subjects made this choice, which led Cosmides to the conclusion that there is no sixth sense for altruism. Again, note the form of the argument: a connection is made between on the one hand a supposed generically determined module which would be capable of determining whether a person is willing to pay a price without claiming the corresponding benefit, and on the other hand the lexical element ‘altruist’. It is rather questionable whether there exists such a close connection between genes and expressions in natural language. These concerns become pressing when they involve loaded concepts such as ‘aggression’: if someone claims to have discovered a ‘gene for aggression’, what has actually been discovered?

Let us consider first the definition of ‘altruism’ as given by the Oxford English Dictionary (OED): “Devotion to the welfare of others, regard for others, as a principle of action; opposed to egoism or selfishness.” Is it possible to transform this definition into a prediction about behaviour, and if so, is this prediction consistent with that of Cosmides?

Cosmides briefly toyed with the possibility that the Stanford University undergraduates she used as subjects did not know the meaning of the word ‘altruism’, and she therefore performed another experiment in which the word ‘altruistically’ was substituted by ‘selfless’, a substitution which led to somewhat larger number of ‘competence’ answers, although not significantly. She concludes that the problem cannot reside in the meaning of ‘altruism’, and hence that reasoning with social contracts, in so far as it concerns the detection of algorithms, is just as abysmal as reasoning with abstract conditionals. This conclusion points to a tin ear for semantics: is it really plausible that a many-faceted concept such as ‘altruism’ can be captured in a few simple logical rules?

We can see, for instance, that the story about Big Kiku’s dealings suggests the opposite of altruism; a truly altruistic person would give that cassava

root, no questions asked, without demands. It is true that Cosmides has done another experiment, where the preamble to the main question now presents Big Kiku as generous (which resulted in roughly the same scores), but this does not obviate the main problem: that a conditional promise is not altruistic if we follow OED's definition; only an unconditional promise would count as such. If this is so, then *no* card needs to be turned—one can see immediately that Big Kiku is not an altruist.

One arrives at the same conclusion if one argues as follows: 'The cards exhibited make plain that at least one of the men did not get cassava. That's not altruistic: a true altruist feeds the hungry. Hence one doesn't have to turn a card to see that Big Kiku is not an altruist.' We now have two different predictions, Cosmides', and the 'no card' prediction.

A different prediction from these two can be obtained when the subject argues as follows: 'Big Kiku has done a conditional promise. Altruism requires that Big Kiku at the very least keeps his promise, but furthermore displays his generosity. But then all four cards have to be turned.' Hence we have at least three different predictions.

It is therefore unclear what follows from the supposed existence of an altruism detector, and hence what counts as a falsification. Furthermore, subjects may experience great confusion due to the several possibilities, which means that they may become even more susceptible to the confusion caused linguistic ambiguities.

Indeed, if we pay close attention to the linguistic form of the instructions, we encounter another problem. In Cosmides' first experiment subjects were asked to determine whether 'Big Kiku has broken his word to any of these four men', after being asked: 'Did Big Kiku get away with cheating any of these four men?' It so happens that there exists an extensive linguistic literature on the word 'any' (technically, a determiner), which is agreed upon the view that 'any' can have two senses, as evidenced by the sentences 'if there's anything I can do for you, tell me' en 'anyone with an ongoing flight is requested to go the transfer desk'. In the latter case, 'anyone' means 'everyone', but in the first case it means 'something (whatever)'. In logical terms: 'any' can be both a universal and an existential quantifier.

One can see why Cosmides wanted to use 'any': this word is often used to decrease the set of legitimate exceptions. Often, superficially universal statements allow for self-evident exceptions, which makes the following sentence acceptable: 'Every match I strike lights ... Not *any* match of course, a wet one doesn't.' The italicised *any* expresses that this determiner is emphasised; precisely because wet matches are not considered, is the sentence 'Any match I strike lights' false in this context. Rules or contracts as used

by Cosmides may have such self-evident exceptions. A conditional promise such as ‘If you get a tattoo, I’ll give you cassava root’ may have all kinds of silent annulling conditions, like ‘but obviously not if you betrayed my brother to the enemy’. Unfortunately the introduction of ‘any’ generates as many problems as it solves.

Consider the occurrence of the word ‘any’ in the question: ‘Did Big Kiku behave altruistically towards any of these four men?’ It is generally assumed that ‘any’ here means ‘somebody (whoever)’, and *furthermore that the implied answer to the question is negative*, at least when ‘any’ is emphasised. In the experiment as performed by Cosmides one cannot control for subjects’ mental representation of ‘any’: with or without emphasis. Let’s see how the meaning of ‘any’ influences the processing of the experimental instructions. First: ‘Did Big Kiku get away with cheating any of these four men?’ (we disregard the needless complications introduced by the expression ‘get away with’). On the one hand, since ‘any’ in this question means ‘someone’, it seems as if the problems with dependencies between card choices resurface here: if it suffices to find *one* person who has been cheated, then the subject may argue as follows: ‘I should first turn the ‘got the tattoo’ card, and, dependent upon the outcome of this experiment, possibly also the ‘Big Kiku gave nothing’ card.’ Cosmides’ first experiment would thus have something in common with standard experiments on the abstract four card task. However, if it is true that the expected answer is negative<sup>1</sup>, and hence that the question suggests that no one has been deceived, then the subject is led to turn *both* the ‘got the tattoo’ and the ‘Big Kiku gave nothing’ cards; it is immediately clear that the other cards cannot be examples of cheating. And indeed this was the most common response.

The situation is more complicated still in the case of the ‘altruism’-experiment, where the instruction is to answer the question: ‘Did Big Kiku behave altruistically towards any of these four men?’ If the above observation about the pragmatics of ‘any’ is correct, the expected answer is negative<sup>2</sup>. But in this case no card shows immediately, i.e. without turning, that Big Kiku has not behaved altruistically toward a particular person. In fact, when deriving possible predictions above, we implicitly showed that for any card there are arguments *pro* and *con* including that card in the selected subset. Thus subjects may get mired in confusion: if ‘any’ expects a negative answer, then the ambiguities surrounding ‘altruism’ do not favour

---

<sup>1</sup>Even if that is true, the pragmatics may well be in conflict with the setting of the story.

<sup>2</sup>In this case the pragmatics is more in line with the setting of the story.

a unique subset, and if 'any' is taken to mean 'someone' then dependency problems resurface.

Then there is the possibility that the question 'Did Big Kiku behave altruistically towards any of these four men?' is taken to mean: did Big Kiku behave altruistically toward this group of four? In that case another problem comes to the fore, the possibility of exceptions. As we have seen, even in the case of the abstract rule, subjects count with the possibility of exceptions. In the case of dispositions such as 'being altruistic' common sense automatically takes exceptions into account. What if Big Kiku refused cassava to one man, but not to the remaining three? It is totally unclear what to predict; perhaps that all cards should be chosen. Here it is very unfortunate that Cosmides reports her data in such a form that only 'successes' and 'failures' are tabulated, i.e. conformity or non-conformity with her predictions. It would have been very informative to also see the pattern of the 'failures'.

## Chapter 7

# Logic and the brain

Nowadays almost all scientists believe that the execution of any cognitive function is dependent upon brain activity. (There have been prominent exceptions to this ‘materialist’ view in the recent past, for example Gödel.) But even so, the relevance of knowledge about the brain to the study of cognition is disputed. Here are two contrasting quotes:

I shall therefore assume without scruple at the outset that the uniform correlation of brain-states with mind-states is a law of nature. ... To some readers such an assumption will seem like the most unjustifiable a priori materialism. ... But although we affirm that the coming to pass of thought is a consequence of mechanical laws ... we do not in the least explain the nature of thought by affirming this dependence. (William James [54, p. 6])

The idea that ultimately there should be a unified theory of the brain—a theory that encompasses all levels of description—has of course been around for a long time. But the idea has typically seemed both surprisingly vague and pathetically remote ... But things are changing. Developments in neuroscience and in philosophy, as well as developments in psychology and computer science, have brought the disciplines to a stage where there are common problems, and there is a gathering sense of the benefits for cross talk ... We have entered a time when the idea of a unified theory of how the brain works is no longer impossibly remote. (Patricia Smith Churchland [16, p. 5-6])

The quote from James is from 1892, but until recently his view was dominant. The advent of powerful brain-imaging techniques has changed the picture somewhat. It is questionable whether this is due to an increase in knowledge or to the change in perspective that inevitably accompanies ‘big science’. The purpose of this chapter is to outline what is known about the instantiation of logic in the brain, and, more importantly, to discuss what *kind* of insight can be obtained by means of these techniques.

## 7.1 Avenues to brain-correlates of reasoning

Much research is concerned with the attempt to find an approximate location in the brain for a particular cognitive function. The paradigmatic example is vision, where a number of areas have been identified which have definite functions, such as the detection of the orientation of a line, color, or the detection of motion. These functions can be characterised by relatively simple and well-understood deterministic algorithms, and hence it is not very surprising that the functions are subserved by ensembles of dedicated neurons. Furthermore it is extremely important that in the case of vision the stimulus can be controlled completely, so that one has a fairly good idea of what a subject is doing.

Language also provides an interesting example. Although the claim is not uncontroversial, many scientists now believe that there exists a ‘language organ’, a part of the brain dedicated exclusively to syntactic processing. A recent paper paints the current picture as follows:

Syntactic abilities are distinct from other cognitive skills, and [are] represented entirely and exclusively in the left cerebral hemisphere. Although more widespread in the left hemisphere than previously thought, they are clearly distinct from other human combinatorial and intellectual abilities . . . language is a distinct, modularly organized neurological entity. Combinatorial aspects of the language faculty reside in the human left cerebral hemisphere, but only the transformational component (or algorithms that implement it in use) is located in or around Broca’s area. (From the abstract of Grodzinsky [40].)

In general the goal here is to find particular areas of the brain that perform a given syntactic computation.

It is much less clear what to expect in the case of cognitive functions such as reasoning, that do not seem to correspond to unique deterministic

algorithms<sup>1</sup>. As we have seen, the evolutionary psychologists claim that successful reasoning can only be due to the presence of an evolved module (such as cheater detection), and if this view were correct, one might be led to expect a well-defined ‘seat of reason’. But the evolutionary psychologists’ view turned out to be much too simplistic. Issues of truth, planning, induction and meaning were shown to play a role in overt reasoning on the selection task, and hence one can plausibly expect a distributed pattern only. The reader may be tempted to add that for instance vision should also be added to this list, since the argument to be judged, say a syllogism, is usually presented visually. However, the design of a brain-imaging experiment tries to factor out this component, by presenting the argument twice and subtracting the data obtained. That is, the first time the argument is presented the subject is asked to give a judgement about the meaning of the premises, for example whether the premises make sense. Recording a subject’s brain-activity while he gives the judgement, gives what is called the *baseline* which supposedly contains all the activity pertaining to the visual and language processing aspects of the task. The argument is then presented a second time, and now the question is whether the conclusion ‘logically follows’, or ‘necessarily follows’ from the premises: the *deduction condition*. Subtracting baseline from deduction condition should then give the essential reasoning component of the task. Clearly this methodology only works if the stages of interpretation and reasoning can be cleanly separated. We have seen some evidence that this is not always so. In any case it is necessary to assume that a subject’s reasoning process is not activated before he has been told explicitly to reason. This is not yet to say that the assumption is not warranted in the fairly simple reasoning processes studied in the brain-scan literature, but it shows that the methodology may not extend to the more complex cases.

Much of the literature applying brain scans to the study of reasoning is concerned with the supposedly mutually exclusive theories ‘mental logic’ and ‘mental models’<sup>2</sup>. Here is a representative quote

---

<sup>1</sup>For a recent counterblast to the idea that brain-imaging can lead to identifiable correlates for higher cognitive functions, see Uttal [105]. He claims that localisation has never been well demonstrated for any cognitive process. Furthermore there are a number of reasons why the attempt to localise higher cognitive functions must fail: brain regions are not sharply demarcated, regions generally have intricate interconnections, and lesions establish necessity but not sufficiency.

<sup>2</sup>The papers Stenning and Oberlander [97] and Stenning and Yule [99] contain results which make it rather doubtful whether these theories in their present incarnation can be distinguished at all.

[Mental logic claims] that deductive reasoning is a rule governed syntactic process where internal representations preserve structural properties of linguistic strings in which the premises are stated. This linguistic hypothesis predicts that the neuroanatomical mechanisms of language (syntactic) processing underwrite human reasoning processes . . .

[Mental models claims] that deductive reasoning is a process requiring spatial manipulation and search. Mental model theory is often referred to as a spatial hypothesis and predicts that the neural structures for visuo-spatial processing contribute the basic representational building blocks used for logical reasoning. (Goel et al.[34])

It then seems possible to set up an experiment to distinguish between the two theories, by using experimental evidence that language processing is localised in the left hemisphere, and spatial processing in the parietal cortex. If a subject shows no activation in the parietal cortex while reasoning<sup>3</sup> this is taken to be evidence against the involvement of spatial processing, and hence ultimately against mental models. A result of this kind was indeed obtained by Goel and colleagues [35].

We have no particular desire to defend mental models here, but it should be said that the logic of the argument leaves something to be desired. It is a commonplace observation that the derivation of a prediction makes use of auxiliary theories, and that a test of the prediction tests the auxiliary theories as well. But if one is somehow convinced that logical reasoning also involves spatial processing, why not conclude that spatial processing is apparently not restricted to the parietal cortex? Or that semantic processing involves spatial processing as well (i.e. also before the subject engages upon the reasoning task), so that spatial processing largely becomes invisible after subtraction? The latter hypothesis is not implausible if it is indeed true that interpretation and reasoning cannot be separated completely.

Also the statement that mental logic would predict left hemisphere activation may need qualification. The argument given is that ‘This linguistic hypothesis predicts that the neuroanatomical mechanisms of language (syntactic) processing underwrite human reasoning processes’, but if it is true

---

<sup>3</sup>These experiments are usually done on a number of subjects, say 10, in an attempt to cancel out individual differences. Results are therefore reported in the form: ‘no activity up to significance level  $p$ ’. This is then taken to imply that, for this particular task, for most humans, for most of the time there is no activity in the area considered. This conclusion would be more compelling if backed up by longitudinal studies.

that the 'language organ' is a modular entity distinct from other combinatorial capabilities (cf. the quotation from Grodzinsky above), then there does not seem to be much reason to expect proximity of language and reasoning in the brain.

We will now consider some of the ways in which brain correlates of reasoning have been investigated.

**Lesions** These can be local or global, stable or progressive. For the study of brain correlates of cognition, local and stable lesions are of greatest importance. Examples are head-injuries (due to accidents) and strokes. These often knock out particular areas of the brain, and if this is accompanied by a loss of a function, we may learn something about the location of that function. A famous example is the effect of the loss of the amygdala, relay station on the fast visual tract from retina to motor cortex. The amygdala is believed to provide very coarse visual processing, in order to quickly alert the motor system to the presence of possibly living things. (Along a parallel route, slower but finer visual processing is going on, whose end result may inhibit the motor process started by the amygdala.) It has been observed that the loss of the amygdala leads to a narrowly circumscribed semantic deficit, in which the impaired subject is unable to retrieve names of living things. It is tempting to conclude that therefore these names are stored in the amygdala, but other conclusions are possible, for instance that a critical path has been interrupted.

The lesions discussed are examples of 'natural' lesions. Sometimes lesions are due to surgical intervention, as when the connection between the two hemispheres is cut, an intervention necessary to treat severe cases of epilepsy. In this case we may obtain information about the involvement of the hemispheres in a particular cognitive function. For instance, one may present a stimulus to the right visual field, so that it will be processed by the left hemisphere only. If the connection between the hemispheres is cut, the right hemisphere will not be activated, and one may then see whether processing of the stimulus is different when compared to the case where both hemispheres are activated.

**Electroconvulsive therapy (ECT)** This treatment of mental disease can be informative about cognitive function when it is applied hemilaterally. In this case, only one hemisphere is suppressed temporarily (30 to 40

minutes)<sup>4</sup>, leaving the other hemisphere in a higher state of activation. One may thus obtain information about how the hemispheres process a stimulus. As in the previous case, it is impossible to do controlled experiments here; so one cannot control for the effect of the mental disease upon the processing.

### 7.1.1 Positron emission tomography (PET)

[to be supplied]

### 7.1.2 Functional magnetic resonance imaging (fMRI)

[to be supplied]

## 7.2 Experimental results

### 7.2.1 Unilateral brain lesions and the Wason selection task

Golding [36] investigated the performance of subjects who had lesions in either the left or the right hemisphere<sup>5</sup>, on the Wason selection task. The reason why such an experiment could possibly be interesting, is that part of the difficulty of the task may be due to interference between the two hemispheres; as Golding writes

It was postulated that visual skills known to be lateralised to the right hemisphere inhibited the verbal skills of inference, thought to be lateralised to the right hemisphere, thus preventing insight into the problem.

She therefore predicted that patients with right hemisphere lesions would perform better on the selection task, and that this success would be related to some deficit in perceptual classification. The visual skills that interested her particularly have to do with seeing objects from unconventional angles. For instance, some patients with right parietal lesions have difficulty recognising a picture of a bucket when it has been photographed from above, whereas they experience no such difficulty when the bucket has been photographed from the side; in fact these patients positively deny that the former picture represents a bucket even when told so. (This phenomenon was first observed by Warrington and Taylor [108].) In this sense these patients

---

<sup>4</sup>It is not quite clear what ‘suppression’ means here, since subjects whose left hemisphere has been ‘suppressed’ are still able to engage in conversation.

<sup>5</sup>The paper cited does not give details about the nature of the lesions.

have a problem with perceptual classification. It is not quite straightforward to see what might cause this difficulty. The theory put forward by David Marr ([73, p. 328]) is that objects come with a natural coordinate system, and that in a unconventional view the main axis is foreshortened.

The pattern of the results obtained by Golding is striking. Of her subjects with right hemisphere lesions (RHL) 30% made the correct  $p, \neg q$  choice, 40% chose  $p, q$ , 20% chose  $p, q, \neg q$ , nobody chose  $p$ , and 10% miscellaneous. In the LHL group these figures were 5%, 15%, 0%, 50% and 20%. The controls<sup>6</sup> showed the usual pattern: no correct answers, 55%  $p, q$ , 30%  $p$ , all else miscellaneous. Furthermore, and more striking still, of the 10 RHL subjects who chose either  $p, \neg$  or  $p, q, \neg q$ , 9 showed a deficit on the unconventional angles test, and of the remaining 10 who made other choices, none had any deficit on this test.

The paper [36] does not really offer an explanation beyond saying that the visual coding of the task apparently interferes with the verbal processing. The interference goes both ways: the cards mentioned in the rule would dictate where visual attention is directed, which in turn would determine which card would be considered relevant for logical reasoning. In contrast, patients who made either the  $p, \neg q$  or  $p, q, \neg q$  choices referred to all cards in their deliberations, which may be taken as evidence that their visual attention was not constrained by the cards mentioned in the rule. But what is the role of the very specific unconventional views deficit?

### 7.2.2 ECT and syllogistic reasoning

Further confirmation of the role of the right hemisphere in reasoning was obtained by an ECT experiment performed by Deglin and Kinsbourne [21]. Manic-depressive and schizophrenic patients were given arguments of the form  $\forall x(Ax \rightarrow Bx), Aa/?Ba?$ <sup>7</sup> to solve, after having received ECT to one of the hemispheres; sometimes the premises were familiar, sometimes unfamiliar and sometimes plainly false. Here are some examples:

[familiar] All rivers contain fish.

The Neva is a river.

?Does the Neva contain fish?

[unfamiliar] All countries have flags.

---

<sup>6</sup>The average age in this group was much higher than usual, since they were matched to the average age of the RHL's and LHL's.

<sup>7</sup>This type of material is customarily called a 'syllogism', although it is not a syllogism in the Aristotelian sense. We shall follow custom here.

Zambia is a country.  
 ?Does Zambia have a flag?

[false] All countries have flags.  
 Quetzal is a country  
 ?Does Quetzal have a flag?

The pattern of results was consistent. When the right hemisphere is suppressed, subjects calmly accepted the premises and affirmed that the conclusion follows, sometimes showing irritation at being asked such a simple question. When the left hemisphere was suppressed<sup>8</sup> however, the experimenters obtained answers such as

There used to be lots of fish in the Neva, but the river is now so polluted it is completely dead!

Is there really such a state, Zambia? Where is it? Who lives there?

How can I know? I don't even know that country [i.e. Quetzal]!

The authors conclude from this that the left hemisphere is capable of performing formal logical operations independent of truthvalue; in contrast, the right hemisphere 'seems incapable of the willing suspension of disbelief' and in any case unable to abstract from truthvalue. Note the similarity of the answers of the LH suppressed patients to those given by subjects in Luria's and Scribner's experiments, cf. section 1.6.

It is also of some interest to relate this pattern of results to those obtained by Golding and others on the selection task. The similarity is that suboptimal functioning of the right hemisphere facilitates logical reasoning. If it is correct that an active right hemisphere makes it more difficult to abstract from truthvalue, then one can begin to see why some subjects take the rule in the selection task to be true, instead of investigating its truthvalue. Even so, it remains somewhat mysterious why the unconventional views deficit should be a determinant of good performance.

### 7.2.3 Brain-imaging studies of deduction

As remarked above, the few brain-imaging studies that have been done, have concentrated on the supposed dichotomy between mental logic and mental

---

<sup>8</sup>Since subjects could still comprehend natural language, the suppression cannot have been total.

models theories. If mental logic is right, this would lead to left hemisphere activation; if mental models is right, this would lead to right hemisphere activation, in particular that of the right parietal cortex. Right hemisphere activation of the prefrontal cortex, the parietal cortex and the occipital cortex has been linked to spatial working memory tasks. Furthermore, the parietal cortex is involved in the encoding an object's position with respect to landmarks in the 'where' system of the brain. It was therefore thought plausible that, if mental models is correct, at least some degree of right hemisphere activation should be noticeable. Even before we come to the experimental results, we must note that this prediction needs to be qualified: we have seen that mental models for syllogistic tasks incorporate various abstraction devices which have nothing spatial *per se*. If mental models (including abstraction devices) is a good paradigm, one should therefore expect also left hemisphere activation.

In addition to standard syllogistic material, Goel et al. [35] added examples of relational reasoning involving both spatial and nonspatial predicates. Examples are

[standard syllogism] Some officers are generals.  
 No privates are generals.  
 Some officers are not privates.

[spatial relational] Officers are standing next to generals.  
 Privates are standing behind generals.  
 Privates are standing behind officers.

[nonspatial relational] Officers are heavier than generals.  
 Generals are heavier than privates.  
 Privates are lighter than officers.

The reason for including the spatial relational condition is that

Perhaps if the argument contents [of previous experiments] had explicitly required spatial encoding, they may have found right hemisphere and parietal encoding. We did not consider this a very strong possibility because both introspection and some cognitive theories suggest that syllogisms are mapped onto spatial representations by way of Venn diagrams or Euler circles. But it seemed necessary . . . to test this hypothesis.

The subjects in the experiment first had to undergo a test of their spatial abilities. It was hypothesised that subjects with high spatial ability would

show more parietal activation. It turned out that for both relational conditions, the areas activated were the same, and nonparietal. The pattern for the syllogistic condition was slightly different, but also 'nonspatial'. Interestingly, there was no significant difference between the 'high spatial' and 'low spatial' groups.

What to conclude from this? The authors cite a paper from Kosslyn et al. [65] which reports a dissociation between categorical spatial encoding (i.e. of spatial relationships as for instance embodied in natural language), and coordinate spatial encoding. The former is localised on the left, the latter on the right. The authors draw the conclusion from all this that 'spatial relations in linguistic reasoning bypass the parietal system and are encoded directly into the language system'. They comment

[This hypothesis] is counterintuitive. Most of us have the phenomenological experience of visualizing syllogisms as Euler circles or Venn diagrams or mapping spatial relational arguments onto a spatial matrix when solving these problems. It is possible that this experience is epiphenomenal and that the real work is being done by very different mechanisms.

This conclusion hinges on the assumption that spatial processing is exclusively localised to the right parietal lobe. But why stick to this assumption, in the face of evidence that some reasoning with qualitative spatial relations occurs in the left hemisphere? If mental models theory is cast in such a form that it somehow involves translation of sentences into qualitative spatial relationships, then the brain-imaging results do not decide between mental logic and mental models.

## Chapter 8

# Autism, development and evolution

Autism is an abnormality of human development which has been argued to throw light on human development and evolution, and specifically on the evolution of mind. We explore it here for the light it can throw on several cognitive science questions—particularly the relation of pragmatics to language; the modularity of the mind; the relation of reasoning to emotion; and the relation between simulation and theory in the computational architecture of mind.

Autism is a clinical syndrome first described by Leo Kanner in the 1950s, often first diagnosed in children around 2–3 years of age as a deficit in their affective relations and communication. The autistic child typically refuses eye-contact, is indifferent or hostile to demonstrations of affection, and exhibits delayed or abnormal communication, repetitive movements (often self-harming) and is indifferent to pain. Autistic children do not engage spontaneously in make-believe and show little interest in the competitive social hierarchy, and in personal possessions.

Autism comes in all severities—from complete lack of language and severe retardation, to mild forms continuous with the ‘normal’ range of personalities/IQs. Autism is sometime distinguished from Asperger’s syndrome—‘autism without language impairment’?—but Asperger’s is probably just the mild end of the autistic spectrum. Autistic children share many symptoms shown by deaf and by blind infants.

There are known biochemical abnormalities associated with autism. There is some indications of a genetic basis. Psychological analyses of autistic functioning are not inconsistent with or exclusive of such biochemical or genetic

level analyses.

One of the laboratory paradigms that has been shown to have some success in distinguishing autistic children from controls matched for verbal intelligence is what is known as the false belief task (Perner *et al.* [79]). A child shares with a doll some information about where an object starts off—say a sweet hidden in a drawer. The doll then goes off-stage and the sweet is moved to another hidden location—say a box. The doll now re-enters and the child is asked where the doll will look for the sweet. Before the critical point in development (about 3.5 years) the normal child says that the doll will look where the child knows the object is; after that watershed, the child predicts that the doll will look where the doll falsely believes the object is. Autistic children tend to continue failing this task after normal children, or even children of matched verbal intelligence, pass it.<sup>1</sup> More complicated iterated false belief tasks catch less severely autistic children and still can differentiate them from controls.

There are many questions about what ability is being measured here. For example, Peterson and Riggs [80] show that failure on the false-belief tasks are highly correlated with failure on inferentially closely related counterfactual tasks. The child observes the doll move the hidden object from the cupboard to the refrigerator in the course of baking a cake, and is then asked where the object *would be* if the doll had *not* baked a cake. Failure to place it in the cupboard is failure at counterfactual reasoning about what *would* have been the case.

So how are we to characterize what failure is going on here—lack of ability to reason about beliefs (especially false beliefs), or lack of ability to reason counterfactually? False-belief tasks involve a kind of counterfactual reasoning, though counterfactual reasoning is a much broader category than false-belief reasoning. This is a typical example of the problems of specifying *what* is involved in the capacities which are required by these tasks.

There are several current psychological theories of autism. To mention some of the main ones: the theory-of-mind deficit theory (Leslie [67]); the affective foundation theory (Hobson [50]); the executive function deficit theory (Russell [84]); the central coherence theory (Happé [42]).

The ‘theory theory’ originated in Premack’s work on non-human primates attempting to characterise the differences between human and non-human primates. Alan Leslie ([67] proposed that human beings have a brain

---

<sup>1</sup>One might wonder how the autistic child, who is deemed to have problems with entertaining phantasies, is nevertheless supposed to be able to imaginatively construe the doll’s participation in this experiment. On the other hand, there is evidence autists can engage in phantasy when instructed to do so.

‘module’ that does reasoning about minds, by implementing a ‘theory of mind’, and that autistic development could be seen as delay or impairment in this module. So, the theory theory goes, in normals the module constitutes the difference between humans and their ancestors. To paint with a broad brush, this claim gives us the two equations:

$$\begin{aligned} \text{chimpanzee} + \text{ToM} &= \text{human} \\ &\text{and} \\ \text{human} - \text{ToM} &= \text{autist} \end{aligned}$$

Certainly this is a great simplification, but we will argue below, a useful one. These equations raise many questions about just what non-human primates can and can’t do in the way of reasoning about conspecifics behaviour and mental processes. Apes have considerable facility in reasoning about conspecifics’ *plans*—the intentions behind their behaviour, but they appear not to be able to reason about conspecific’s epistemic states.

A related point is that young children seem to develop ‘desire’ psychology before ‘belief’ psychology, although desires are ‘states of mind’, and so it is unclear why they do not require a theory of mind to reason about them? This is the entry point for Harris’ alternative simulationist account. More later.

The theory-theory can be taken as explanatory, or as an important label for a set of problems. Its authors appear to claim the former. It seems more plausible at this stage to interpret it as the latter—an important label for a problem.

Hobson’s theory of autism does not so much *deny* Theory of Mind (henceforth ToM) as seek to *derive* it from more fundamental ontogenetic processes—in particular from the affective foundations of interpersonal communication. Human uniquely control shared attention, especially by gaze. ‘Inter-subjectivity’ is established through mutual control of attention. Just as Piaget saw the child’s sensorimotor activity as achieving the child’s mastery of where itself left off and the world began, so Hobson sees the child’s understanding of itself as a social being separated from others being achieved through joint attentional activity. The child must learn that the other can have different representations, and different values. Hobson proposes that it is autists’ *valuation* of these experiences of intersubjectivity which is abnormal. If the child does not experience the achievement of intersubjectivity as rewarding (or even experiences it as aversive), then any cognitive developments founded on it will not develop normally. Cognitive symptoms of autism are, on this theory, *consequences* of this valuation.

Russell's executive function deficit theory is built on the observation that autists often exhibit severe perseveration—they go on carrying out some routine when the routine is no longer appropriate, and exhibit great difficulty in switching tasks. This perseveration is observed in certain kinds of patients with frontal cortex damage, and, so the theory goes, gives rise to many of the symptoms of autism: obsessiveness, insensitivity to context, inappropriateness of behaviour, literalness of carrying out instructions. Although autists lack spontaneity, they may be able to carry out tasks when instructed. Phantasy play is an example.

Happé's central coherence theory of autism is built on the observation that autists show certain *supernormal* abilities, particularly on some visual tasks (e.g. Hidden Figures), and *lack* of susceptibility to some visual illusions (e.g. Muller-Lyer). Autists are good at things which can be done by attention to detail while ignoring 'the big picture'. An example of the hidden figures test is presented in Figure 8.1. We return to this test when considering learning styles exhibited in the normal population.

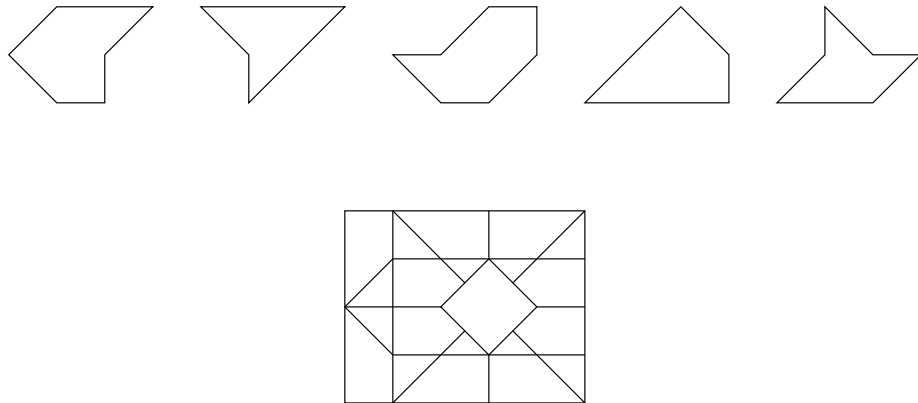


Figure 8.1: Example Hidden Figures Test item. The task is to select which of the outline figures in the row above is contained in the complex pattern below.

This quick survey of theories is intended to illustrate that there are multiple psychological characterisations of autism. Although it is the habit of the field to present these theories as alternatives, it is far from obvious that they are mutually exclusive. Hobson's theory might be construed as a theory about where our theory-of-mind abilities develop from, and although there is some tension between the idea that they come in a genetically determined

module, and that they develop out of experiences of communication, in truth these are not really inconsistent, and anyway, the theory theory does not present any evidence for genetic basis or modularity save for the lack of any evidence of learning and the discreteness of the diagnosing task. Our ability to walk is ‘innate’ but develops out of experiences. Similarly, executive function and central coherence are presumably computational properties of systems which might or might not be components of whatever system realises ‘theory of mind’ abilities.

What these theories do is raise acute theoretical questions about just what their terms of art mean ‘module’, ‘theory’, ‘innate’ etc. The authors do not look for guidance to other disciplines whose terms these are. In computer science, when a module is proposed, this immediately invokes issues about how the module functions with other modules (especially how it communicates with other modules) so that the system can achieve its ends. In the rhetoric of psychology, ‘module’ seems to function mainly to negate some supposed theory that the mind is homogeneous, but does not seem to be followed up by questions about communications between modules. Similarly, modern genetics offers very sophisticated understandings of interactions of genes and environment in the determination of phenotype. A percentage of variance in a trait controlled by genes can be calculated, perhaps giving a truly precise meaning to ‘innateness’. But this can only be done *relative to an environment*. There are well known cases of traits 100% controlled by genes in one environment and 0% controlled by genes in another.<sup>2</sup> The philosophy of science gives us a sophisticated analysis of what theories are (and are not). Theories are *not* (simply) sets of sentences reasoned over by a theorem prover, but are also grounded by practices of empirical observation and practices of social validation. Psychologists have not asked what they mean by a ‘theory of mind’ embedded inside the mind of a ‘theorist’. Is this intended to specify some computational architecture, or is it just a name for whatever ‘magic’ does the business? And just what business does it do? Beyond passing the false-belief test?

The usual situation in psychology is that the ‘theories’ or proposals or hypotheses are not well enough developed to differentiate their compatibilities and incompatibilities very far. This is not to denigrate psychology—understanding the mind is very hard. But it does caution against taking their claims to explain phenomena (rather than to provide useful preliminary labels) without some healthy skepticism.

---

<sup>2</sup>The classic example is wing-veining in *Drosophila* which is genetically controlled in high humidity but environmentally controlled at low humidity.

**What can we expect cognitive science might contribute? And learn?** Cognitive science brings together many bodies of knowledge relevant to understanding autism, and understanding autism offers several exciting possibilities for cognitive science. We will focus on three here:

- place of pragmatics in theories of cognition
- interactions between social, emotional and cognitive phenomena
- theory and simulation in understanding minds

Autistic children's primary deficit is not in language but in communication. Cognitive science (and linguistic) theories often treat the study of language almost disregarding any theory of communication. It talks a 'language faculty' made up of components —phonetics, phonology, morphology, syntax, semantics—but without emphasising the gap between language structure and language function. True, pragmatics now receives considerable attention, but it sits uncomfortably with the other circumscribed components. Autists have relatively unimpaired language structure but their *communication* is profoundly disturbed.

There are many reasons to be cautious about functional explanations of language structure (Chomsky is on record as denying that language has anything particularly to do with communication). But if we are interested in cognition we have no alternative than to formulate theories of language's role(s). Why should language structure ever evolve without *uses*? It may have many uses, and some of its parts may be spandrels, but an account of selection pressures is unavoidable.

This disciplinary fence building often has the consequence that there is an assumption that the mind first processes the structure of language tokens, and then works out how they apply in context. This is an outrageous inference. As we have seen in the selection task, the pragmatic decision that someone intends an utterance to be deontically interpreted may completely override the semantic structure which would be assigned to a sentence out of context.

But autism does not merely challenge cognitive science to develop theories of communication. It also challenges it to develop theories of how the cognitive and the social phenomena of communication fit together. Cognitive and social theories of communication have been developed rather separately so far. 'Propositional' or 'ideational' communication is regarded as a cognitive phenomenon whereas 'phatic' communication is regarded as a

social phenomenon.<sup>3</sup>.

But we communicate languages, not just propositions. That is to say, the result of communication is the establishment of mutually consistently interpreted languages (and sub-languages), and speaking common languages is one of the most powerful definers of social groupings. Every communicative act has propositional and phatic consequences, even if one or the other kind of consequence may be attenuated in many cases. This is perhaps most easily illustrated in educational communication. What did you learn when you just learned the word ‘phatic’? You learned what ‘phatic’ means, but you also became a member of a group of people who can use the word. As we shall see, autism challenges us to understand some of these relations more deeply.

The third issue on which cognitive science has much to offer to the understanding of autism is the issue of how we (normals) understand minds. There has been an extensive debate in recent times between those who believe that our understanding of minds (others’ as well as our own) is based on a theory-of-mind, and those who believe it is based on simulation (much of this debate has appeared in the journal *Mind and Language*). If the theory-theory is to be more than a label for whatever apparatus we have for understanding minds, then psychology needs to specify what the theory theory means a great deal more precisely.

Let us start by contrasting the idea of simulation with theory. How can we predict what a tailless aeroplane will do? One way would be to develop a theory of aerodynamics which was based on laws and allowed predictions to be made about cases such as this. Another quite different approach would be to build a simulation—a model aeroplane in a wind-tunnel perhaps. Simulation does not depend on having an explicit principled theory of flight. We put the model in the wind-tunnel, watch what happens, and draw certain inferences. How does this analogy apply to predicting/understanding minds?

Simulation theorists suggest that people can predict what people will do (themselves included) by ‘setting certain parameters’ and running ‘off-line’ whatever mechanisms it is that allow them to generate behaviour themselves. So if I want to know what Fred will do in a certain situation, I imagine what I will do (perhaps changing some parameters which I know to distinguish myself from Fred) and then run my planning mechanisms and ‘read off’ what they tell me to do.

---

<sup>3</sup>Phatic communication is communication which establishes or maintains social groupings—fashion and religious ritual are two examples

We must have mechanisms for planning our own behaviour even if we do not have access to all the principles involved. Clearly we can also run this mechanism off-line—we can plan without executing behaviour. We don't need theories, and we don't have theories a simulationist might argue. If we did have theories, psychologists would be out of a job.

How well does the aeroplane analogy work for the mind? Simulation works well for the aeroplane case because we knew a lot about the relevant correspondences between model and plane which allow the inferences. Shape is relevant. Colour is not. Texture is intermediate—it doesn't matter too much at the low speeds the early experiments were done at. We did indeed develop simulations of flight before much theory.

But our simulations of minds are computational simulations, more like weather simulations than wind-tunnels. Our minds process information about the behaviour to be planned/predicted/understood. Weather simulations could only be developed because they were based on theories of weather—fluid statics and dynamics etc. etc. This does not argue against simulation generally but does argue that we have to be rather careful about how the analogy is used, and be aware that the relations between simulation and theory may be rather complicated in the case of minds. For example, learning new principles of peoples' behaviour appears to alter our simulations, and it may well be that we can learn principles from studying the output of our simulations under systematically varied parameter settings.

On the other hand, theory theorists have criticised simulation accounts on the basis that we clearly cannot predict our own behaviour, let alone other peoples', and if we used whatever mechanisms we use to produce our own behaviour, then we ought to be able to predict—*ex hypothesi*. For example, Stich has argued from the observation that subjects make incorrect predictions about their behaviour in experimental situations.<sup>4</sup>

But this is surely too harsh a demand of simulations. Things get into the social psychological literature pretty much only if they are counter-intuitive effects. The issue is about what parameters we can and do set in the mechanism we run off-line to plan our behaviour. Perhaps we can't easily 'imagine ourselves' into situations vividly enough to get our off-line mechanism to operate accurately—we don't get into the 'feel' of the situation well enough to realise what we actually do 'on the night'. Perhaps we are too easily influenced in the laboratory situations by rather abstract normative

---

<sup>4</sup>The example taken is the Langer effect in social psychology—that subjects value a lottery ticket chosen themselves above one randomly assigned to them—though they fail to predict they will do this.

models of our behaviour—models of what we feel we ought to do, which may nevertheless still be simulations. Nichols and Stich might reply that whatever our mechanisms are that decide what to do, they actually do work right down to the detail of what actually happens *ex hypothesi*. But that does not mean that we can or must use the whole system when we plan, or that we have only one level of detail at which we can run simulations. It all depends how much of our mechanisms we do (or can) ‘take off line’ and run when we are thinking, and whether the outputs of our simulator are interpreted through our theoretical knowledge to some extent.

Simulations and theories may not be quite so distinct or so independent as at first appears. So just what is the contrast between theory and simulation based understandings of mind? The following are some dimensions of contrast between theories and simulations:

- *explicitness*: theories are bodies of explicit principles: simulations rely on tacit correspondences
- *encapsulation*: theories are encapsulations of knowledge in modules which process independently of other theories: knowledge embodied in simulations is not thereby modularised
- *cognitive penetrability*: theories are penetrable by knowledge relating to that being processed: simulations are cognitively impenetrable
- *personal–subpersonal level*: consciously accessible personal knowledge vs. inaccessible parts of our subpersonal computational machinery
- *propositional–affective*: theories operate on representations of propositions: simulations may include feelings
- *expressiveness*: simulations are, like analogies, inexpressive: theories are expressive

Although the first four dimensions may be important for the purposes of the philosophical debate, it is less clear how helpful they are in a debate about cognitive modules. What does it mean for a principle to be explicit in the workings of a module, if the module is an encapsulated sub-personal unit, as it is assumed to be in the cognitive theories?

Here we will focus on the last two dimensions that have received less attention than the others but seem more cognitively relevant. To take the last dimension first, an important computational property of simulations is that their inputs and outputs are couched in *inexpressive* representation

systems. A weather simulation has to be started by specifying a complete map of temperatures, pressures, wind-speeds, humidities etc. Its output is similarly a map of parameters, or perhaps a fully-ordered sequence of maps. Maps, like diagrams, are inexpressive. They cannot express arbitrary abstractions over worlds. Full-orderings cannot express abstractions over events.

Theories, in contrast, do operate on abstractions. Possibly not arbitrary abstractions, but abstractions nonetheless.

The penultimate dimension (propositional versus affective involvement) is particular to theories/simulations of minds, behaviour and experience. While theories might be able to predict what will be true descriptions of feelings, they will not themselves enact those feelings as part of making predictions or analyses. In contrast, predicting what we (or someone else) will feel in situation X by simulation may involve us experiencing that situation perhaps at some pale level of vividness.<sup>5</sup>

But there is an empirical problem with determining the role of emotional experiences in theory-of-mind inferences. It is obviously too quick to say that theory theories will predict that emotional experiences will not occur during episodes of predicting others' behaviour. We might have a real pencil-paper-and-equations model of mind (imagine yourself as a theorist who finally cracks psychology), and some application of it involving lengthy calculation may lead us to predict some absolutely ghastly experience for an episode's protagonist. At this point we may feel empathy. The empathy was not an implementation of any part of the calculation, but a result of the calculation's prediction. The outputs of the theoretical calculation are fed into our own experiential mechanisms. This is the corresponding difficulty to the possibility of our theories of mind being applied to the outputs of simulations. Once the mind is hybrid, psychology becomes much more difficult. And the mind is hybrid. No one said psychology was easy.

If we could establish conceptual differences between theories and simulations as accounts of minds and how we understand them, then we might be able to empirically test which better accounted for normal and autistic abilities to understand minds. This kind of approach is required to put any

---

<sup>5</sup>This assumption of course treads close to issues about the fundamental possibility of the computational simulation of mind, but we do not need to settle those issues here. We could believe in strong artificial intelligence, and believe that it is possible in principle to give a computational account of mind 'all the way down' but still draw a distinction between how theories of emotion are implemented and how simulations of emotion are implemented. One could even believe that we have both, perhaps one as implementation for routine monitoring and the other for reflective inference.

flesh on either the theory-theory or a simulation theory of mind.

These two dimensions (expressiveness and affective involvement) possibly make two bodies of knowledge applicable to understanding autism. Study of the cognitive impact of diagrammatic as opposed to sentential presentations of information has exploited exactly this issue of expressiveness in empirical studies of learning (Stenning [94]), and shown that there are different learning styles which prefer the two different kinds of presentation. Students who learn well from diagrams can be seen as strategically skilled at using the limited kinds of abstractions expressible in diagrams (or simulations) to solve problems. Students who learn better from linguistic presentations are strategically skilled at exploiting linguistic representations' different mode of expressing abstractions. These learning styles may also offer ways of connecting cognitive preferences with social attitudes to knowledge.

The second dimension that is capable of distinguishing simulation from theory (affective involvement) can be illustrated by the work on conditional reasoning in Wason's selection task described earlier.

Cosmides argued for replacing logical reasoning by what is essentially an emotional reaction—cheating detection. She saw the deontic selection task as being carried out by cheater-detection modules, and the failure of the descriptive task being due to a lack of a module for descriptive reasoning. We, in contrast, argued that the deontic and descriptive tasks posed different problems—the latter a lot more problems. We can use this work to hypothesise a quite different relation between emotion and reasoning.

If we examine Cosmides' 'cheating detection' modules more carefully it becomes puzzling why such a mechanism should not be able to solve the descriptive task. Lying is a form of cheating. Why can the subject not ask themselves whether the source of the descriptive rule could be trying to cheat them into believing a falsehood? Why is this not sufficient to alert them to the not-Q card as evidence of possible cheating?

So the emotional reaction of cheating detection won't explain the difference between the deontic and descriptive tasks. But this line of thought points to a quite different possible involvement of affect in reasoning. It is an interesting psychological hypothesis that at a particular developmental/educational stage, subjects' reasoning might be partly *implemented* in such affective reactions. That is to say, given a scenario which engages their responses to possible cheating or lying, subjects may be able to reason about truth and falsity, where they cannot do so reliably with scenarios which fail to engage these affective responses. After all, subjects' communicative abilities are far more likely to be attuned to understanding speakers' intentions

than sentences' truth values.

At later developmental/education stages subjects may transcend these limitations, and they might even learn to do so by extending their techniques for applying their affective responses in a broader range of situations. For example, these phenomena might underlie some of the effects we observed in the 'truthfulness of unreliable source' condition in the selection task. Many of Piaget's observations of children's reasoning pre-formal operations are quite consistent with such a hypothesis.

However, it is important to note how different this hypothesis is from that entertained by Cosmides and her colleagues. On our hypothesis, cheating/lying detection would be part of what implements both descriptive and deontic conditional reasoning, with their very different logics. On this hypothesis, the relation between representational system and emotion is a relationship between system and implementation. The general human grasp of the different semantics involved in reasoning about how the world *is*, as opposed to about how the world *should* be, remains of paramount psychological importance regardless of implementations. Theorising in terms of emotion does not mean that the need for logical theory disappears.

It goes without saying that on this model emotion and reasoning are conceptualised in a completely different relationship than the conventional tug-of-war between sweet reason and our animal instincts.

**Applying these issues to understanding autism** Researchers in autism might be interested in the learning styles we characterise through learning from diagrams vs. linguistic presentations. Could autism be an extreme learning style? There are some initially suggestive if anecdotal observations. Autism is continuous with personality traits common amongst professions such as engineering which selects for highly visual thinking. Use of an instrument for measuring autistic traits of personality reveals that they are rife in the Cambridge engineering faculty (Baron-Cohen et al. [6]). The 'diagrammatic' learning style we identified is correlated with ability at the Hidden Figures test which autists are actually *better* at than normals.

What our theory has to offer here perhaps is that it is related to a systematic theory of the semantics of diagrams and why they have the impact on reasoning and learning which they are observed to do. We can give an account of the changes in strategies of representation use which happen during learning. It is *not* simply 'visual' vs. 'verbal' thinking. And there are links to be built to social attitudes toward knowledge. Perhaps this prospect of a more articulated theory of learning styles is more important than whether

any particular learning style can be identified with autism.

Even at this early stage, there are puzzles. Autists are supposed to be good at hidden figures because they *can't* focus on the 'big picture'. Our subjects are good at learning from diagrams because they have flexible strategies about when *not* to use them. These sound like opposites. Resolving these conflicts will be a high priority.

What about theory-of-mind vs. simulationist accounts of understanding minds applied to understanding autism? Above we cast some doubt on the idea of theories as *mechanisms*, let alone modular innate mechanisms. We also cast some doubt on how neatly separable theory and simulation are as the basis for understanding minds. Nevertheless, we suggested that if theory and simulation are to be distinguished, then the two most promising dimensions were the inexpressiveness of simulations, and the involvement of affect in understanding. The former thread is what we have just pursued under the idea of autists showing a 'diagrammatic' learning style. The latter seems closely related to Hobson's theories about the developmental origins of autism. If normals understand minds by simulating others' situations on their own 'off-line-planners', and the feelings evoked in this process are causally implicated in the computations, then this would go some way toward explaining why abnormalities of affect are such an important part of autists' abnormalities of understanding minds and communicating with them.

If, in contrast, normals understand each other through a theoretical device which calculates predictions of behaviour from beliefs, desires and intentions of others', then it is far from clear why there should be affective abnormality. It might, of course be, that the abnormality of affect is a secondary consequence of autists' inability to understand others' intentions and behaviour—frustration and fear of an incomprehensible social world are highly understandable. So perhaps the crucial issue here is whether the affective abnormalities associated with autism are primary or secondary.

Harris (e.g. [45]) has done more than anyone else to develop simulationist accounts of normal development, and apply these to autistic behaviour. He sees children's development of mind reading abilities as the growing ability to simulate what minds in different situations than one's own current state will do—the increasing ability to set parameters in one's simulation.

**Concluding thoughts** There is, we believe, something much more general that cognitive science might offer to the field of autism. That something is a contribution resulting from trying to synthesise the many different rel-

evant disciplines' knowledge. It is an alternative pair of equations to those that have motivated interest in autism so far.

At the beginning of this chapter, we found the idea strange that subtracting a ToM from a human being should lead on the one hand to our primate ancestor, and on the other to the autistic. One is the Machiavellian animal: the other a hypo-social being. Perhaps we should consider, at least as a play with ideas, an alternative pair of equations:

$$\begin{aligned} \text{chimpanzee} + \text{magic ingredient} &= \text{human} \\ \text{chimpanzee} + \text{too much magic ingredient} &= \text{autist} \end{aligned}$$

Here the magic ingredient is closely related to what the ToM is supposed to do—take pre-human primate cognition into human cognition, though the mechanism might be simulation instead of theory (or something else entirely). What we are proposing is that autists might be the other side of humans on the evolutionary trajectory.

This is not to deny that autism is pathological. But evolutionary innovations can be pathological. A pathological 'mutation' may be kept in the population by the benefits of a half-dose. Perhaps autism is a double dose? Perhaps autism is the sickle-celled anaemia of human evolution? Sickle-celled anaemia is a pathological genetic condition which in double recessive form leads to adolescent death, but in heterozygous form gives immunity to malaria. It is maintained at levels as high as 10% of populations in sufficiently malarial environments. This is merely a simple genetic case. Autism might have a much more complicated genetics.

What does this mean about the functional nature of the 'magic ingredient'? This proposal probably sounds implausible at face-value. If being able to reason about epistemic states as well as desires is what is novel about human beings, then this is just what autists fail to be able to do in the false belief task. And perhaps there are lots of other differences between chimpanzees and humans, so the equations don't work? All of this may be true, but consider what it is that is grossly different about belief-desire psychology, and desire-psychology *tout court*. Humans can reason about whether epistemic agents are correct about the world, independently of the agents' social position vis a vis the reasoner. They may only be able to do this slightly, and in ideal circumstances, but they can do it at least to some extent. Non-human primates cannot. Autists seem to remove personhood altogether from their ontology. Might this not be construed as overshooting in their effort to factor out the facts from the social hierarchy? This is what one might call the 'holy fool' theory of autism. Autists are seen as having

a cognitive/affective style which removes issues of personal valuation altogether from the equation. Autists just seem a lot more like 'holy fools' than chimps.



# Bibliography

- [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] S. Baron-Cohen and P. Bolton. *Autism: The Facts*. Oxford University Press, 1993.
- [7] J. Barwise and S. Feferman. *Model-theoretic logics*. Perspectives in Mathematical Logic, Springer 1985.
- [8] M. Bowerman. First steps in acquiring conditionals. In [102].
- [9] 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.
- [10] M.D.S. Braine. On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85:1–21, 1978.

- [11] J.D. Bransford, J.R. Banks and J.J. Franks Sentence memory: a constructive versus an interpretive approach. *Cognitive Psychology*, 3:193–209, 1972.
- [12] R.M.J. Byrne. Suppressing valid inferences with conditionals. *Cognition*, 31:61–83, 1989.
- [13] P. Carruthers and P.K. Smith. *Rational ritual: culture, coordination and common knowledge*. Princeton University Press, 1996.
- [14] N. Chater and M. Oaksford. Deontic reasoning, modules and innateness: a second look. *Mind and Language*, 11(2): 191–202, 1996
- [15] K. Cheng, P. and Holyoak, K. Pragmatic reasoning schemas. *Cognitive Psychology*, 14, 1985.
- [16] P.S. Churchland *Neurophilosophy. Toward a unified science of the brain*. MIT Press, Cambridge MA, 1986.
- [17] COBUILD. *Collins Birmingham University International Language Database*. Collins, 1980.
- [18] 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.
- [19] 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.
- [20] D. Cummins. Evidence for the innateness of deontic reasoning. *Mind and Language*, 11: 160–190, 1996.
- [21] V.L. Deglin and M. Kinsbourne Divergent thinking styles of the hemispheres: how syllogisms are solved during transitory hemisphere suppression. *Brain and Cognition* 31: 285-307, 1996.
- [22] K. Dieussaert, W. Schaeken, W. Schroyen, and G. d’Ydewalle. Strategies during complex conditional inferences. *Thinking and reasoning*, 6, 125–160, 2000.
- [23] K. Doets. *From logic to logic programming*. The M.I.T. Press, Cambridge, MA, 1994.

- [24] J.St.B.T. Evans, S.L. Newstead, and R.M. Byrne. *Human reasoning: the psychology of deduction*. Lawrence Erlbaum Associates, Hove, Sussex, 1993.
- [25] 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.
- [26] L. Fiddick, L. Cosmides and J. Tooby. The role of domain-specific representations and inferences in the Wason selection task. *Cognition* 75: 1–79, 2000.
- [27] S.I. Fillenbaum. How to do some things with if. In Cotton and Klatzky, editors, *Semantic functions in cognition*. Lawrence Erlbaum Associates, 1978.
- [28] M. Ford. Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, 54: 1–71, 1995.
- [29] G. Frege. *The Frege reader* (edited by M. Beaney). Blackwell, 1997.
- [30] D.M Gabbay. A general theory of structured consequence relations. In K. Dosen and P. Schroeder-Heister (eds.), *Substructural logics*, Oxford University Press, 1993.
- [31] G. Gebauer and D. Laming. Rational choices in Wason’s selection task. *Psychological Research*, 60:284–293, 1997.
- [32] M. C. Geis and A. M. Zwicky. On invited inferences. *Linguistic Enquiry*, 2:561–566, 1971.
- [33] G. Gigerenzer and K. Hug. Domain-specific reasoning: social contracts, cheating, and perspective change. *Cognition*, 43:127–171, 1992.
- [34] V. Goel, C. Buchel, C.D. Frith and R.J. Dolan Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, 12: 504–514, 2000.
- [35] V. Goel, B. Gold, S. Kapur and S. Houle Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience*, 10: 293–303, 1998.
- [36] E. Golding The effect of unilateral brain lesion on reasoning. *Cortex* 17:31–40, 1981.

- [37] H.P. Grice. Logic and conversation. In P. Cole and J. Morgan (eds.), *Syntax and semantics 3: Speech Acts*. Academic Press, 1975.
- [38] R. A. Griggs. Memory cueing in instructional effects on Wason's selection task. *Current Psychological Research and Review*, 3:3–10, 1984.
- [39] 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.
- [40] Y. Grodzinsky The neurology of syntax: language use without Broca's area. Preprint 2000. To appear in *Behavioural and Brain Sciences*.
- [41] J. Haiman. Conditionals are topics. *Language*, 54:564–589, 1978.
- [42] F. Happ. *Autism: an introduction to psychological theory*. London: UCL Press, 1994.
- [43] P.L. Harris. The work of the imagination. In Whiten, A. (ed.)
- [44] P.L. Harris. *Young children's understanding of pretense*. University of Chicago Press.
- [45] P.L. Harris. *The work of the imagination*. Blackwell, 2001.
- [46] J. Heal. Simulation vs. theory-theory: what is at issue? *Proceedings of the British Academy*, 83: 129–144, 1994.
- [47] J. Heal. Simulation and cognitive penetrability. *Mind and Language*, 11: 44–67, 1996.
- [48] M. Henle. On the relation between logic and thinking. *Psychological Review*, 69:366–378, 1962.
- [49] R.P. Hobson. Against the theory of mind. *British Journal of Developmental Psychology*, 9: 33–51, 1991
- [50] R.P. Hobson. *Autism and the development of mind*. Lawrence Erlbaum Associates, 1993.
- [51] P. Horwich. *Probability and evidence*. Cambridge University Press, Cambridge, 1982.
- [52] E. Husserl. *Logische Untersuchungen. Erster Band*. Volume XVII of *Husserliana*, Nijhoff, 1975.

- [53] E. Husserl. *Briefwechsel*, Volumes I–X. Kluwer, 1994.
- [54] W. James *Psychology, briefer course*. New York, 1892.
- [55] P. N. Johnson-Laird. *Mental models*. Cambridge University Press, 1983.
- [56] P.N. Johnson-Laird and A. Garnham Descriptions and discourse models. *Linguistics and Philosophy*, 3:371–393, 1980.
- [57] P. N. Johnson-Laird and R.M. Byrne. *Deduction*. Lawrence Erlbaum Associates, Hove, Sussex, 1991.
- [58] P.N. Johnson-Laird and P.C. Wason. A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1:134–148, 1970.
- [59] P. N. Johnson-Laird and R.M. Byrne. Conditionals: a theory of meaning, pragmatics and inference. *Psychological Review*, in press.
- [60] P. N. Johnson-Laird and F. Savary. Illusory inferences: a novel class of erroneous deductions. *Cognition* 71(3): 191-229, 1999.
- [61] P. N. Johnson-Laird , P. Legrenzi, V. Girotto, and M. Legrenzi. Illusions in reasoning about consistency. *Science*, 288: 531–532, 2000.
- [62] H. Kamp. A theory of truth and semantic interpretation. In J. Groenendijk, T. Janssen and M. Stokhof (eds.), *Formal methods in the study of language*. Amsterdam: Mathematical Center tracts, 277–322, 1981.
- [63] H. Kamp and U. Reyle. *from discourse to logic*. Kluwer, 1993.
- [64] S. C. Kleene. *Introduction to Metamathematics*. North-Holland, Amsterdam, 1951.
- [65] S.M. Kosslyn, O. Koenig, C.B. Cave, J. Tang and J.D.E. Gabrieli Evidence for two types of spatial representations: hemispheric specialization for categorical and coordinate relations. *Journal of experimental psychology: human perception and performance* 15: 723–735, 1989.
- [66] D. Laming. On the analysis of irrational data selection: a critique of Oaksford and Chater. *Psychological Review*, 103(2):364–373, 1996.
- [67] A. Leslie. Pretence and representation: the origins of a ‘theory of mind’. *Psychological Review*, 94: 412-26, 1987.
- [68] D. Lewis Adverbs of quantification in E. Keenan (ed.), *Formal semantics of natural language* Cambridge University Press, 1975.

- [69] A.R. Luria *Cognitive Development: its social and cultural foundations* Harvard University Press, 1976.
- [70] K. Mani and P.N. Johnson-Laird. The mental representation of spatial descriptions. *Memory and cognition*, 10:181–187, 1982.
- [71] 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.
- [72] H. Margolis *Patterns, Thinking, and Cognition: A Theory of Judgment*
- [73] D. Marr. *Vision* W.H. Freeman, San Francisco, 1982.
- [74] A.Newell and H.A. Simon *Human problem solving*. MIT Press, 1972.
- [75] S. Nichols, S. Stich and A. Leslie. Choice effects and the ineffectiveness of simulation: Response. *Mind and Language*, 10: 437-45, 1995.
- [76] M. R. Oaksford and K. Stenning. Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, memory and cognition*, 18:835–854, 1992.
- [77] M.R. Oaksford and N.C. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101:608–631, 1994.
- [78] M.R. Oaksford and N.C. Chater. Rational explanation of the selection task. *Psychological Review*, 103(2):381–392, 1996.
- [79] J. Perner, S. Leekham and H. Wimmer. Three-year olds' difficulty with false belief: the case for a conceptual deficit. *British Journal of Developmental Psychology*, 5: 125–137, 1987.
- [80] D.M. Peterson and K.J. Riggs. Adaptive modelling and mindreading. *Mind and Language*, 14: 80-112, 1999.
- [81] L.J. Rips. Cognitive processes in propositional reasoning. *Psychological Review*, 90:38–71, 1983.
- [82] L.J. Rips *The psychology of proof*. MIT Press, 1994.
- [83] B. Rumain, J. Connell, and M.D.S. Braine. Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: If is not the biconditional. *Developmental Psychology*, 19:471–481, 1983.

- [84] J. Russell (ed.) *Autism as an executive disorder*. Oxford University Press, 1997.
- [85] S. Scribner *Mind and social practice*. Cambridge University Press, 1997.
- [86] D.B. Skalak and E.L. Rissland. Arguments and cases: an inevitable intertwining. *Artificial Intelligence and Law*, 1:3–44, 1992.
- [87] N.S. Smalley. Evaluating a rule against possible instances. *British Journal of Psychology*, 165(2):293–304, 1974.
- [88] D. Sperber, F. Cara, and V. Girotto. Relevance theory explains the selection task. *Cognition*, 57:31–95, 1995.
- [89] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, 1986.
- [90] K. E. Stanovich and R. F. West. Cognitive ability and variation in the selection task. *Thinking and reasoning*, 4: 193–230, 1998.
- [91] K. E. Stanovich and R. F. West. Individual differences in reasoning: implications for the rationality debate? *Behavioural and Brain Sciences*, 23: 645–726, 2000.
- [92] K. Stenning Anaphora as an approach to pragmatics. In M. Halle, J. Bresnan and G.A. Miller (eds.), *Linguistic theory and psychological reality*. MIT Press, 1978.
- [93] K. Stenning On making models: a study of constructive memory. In T. Myers, K. Brown and B. McGonigle (eds.) *reasoning and discourse processes*. Academic Press, 1986.
- [94] K. Stenning. *Seeing reason: image and language in learning how to think*. Oxford University Press, 2002.
- [95] 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.*, 742–747, 1995.
- [96] 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.

- [97] K. Stenning and J. Oberlander. A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cognitive Science*, 19: 97–140, 1995.
- [98] K. Stenning and M. van Lambalgen. Semantics as a foundation for psychology. A case study of Wason's selection task. *Journal of Logic, Language and Information*, 10: 273–317, 2001.
- [99] K. Stenning and P. Yule. Image and language in human reasoning: a syllogistic illustration. *Cognitive Psychology*, 34: 109–159, 1997.
- [100] K. Stenning and M. van Lambalgen. The natural history of hypotheses about the selection task: towards a philosophy of science for investigating human reasoning. In K. Manktelow and M. Chung (eds.), [Title to be determined.] Psychology Press, 2002.
- [101] M. Tomasello and J. Call *Primate cognition*. Oxford University Press, 1997.
- [102] E. Traugott, A. ter Meulen, J.S. Reilly, and C.A. Ferguson. *On conditionals*. Cambridge University Press, Cambridge, 1986.
- [103] R. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46: 35–57, 1971.
- [104] P. Tulviste. *The cultural-historical development of verbal thinking*. Nova Science Publishers, 1991.
- [105] W.R. Uttal *The new phrenology: the limits of localizing cognitive processes in the brain* MIT Press, Cambridge MA, 2001.
- [106] E.E. Vallduví, E. Engdahl. The linguistic realization of information packaging. *Linguistics*, 34(3):459–519, 1996.
- [107] P. C. van Duyne. Realism and linguistic complexity in reasoning. *British Journal of Psychology*, 65(1):59–67, 1974.
- [108] E.K. Warrington and A.M. Taylor The contribution of the right parietal lobe in object recognition. *Cortex* 3: 152–164, 1973.
- [109] P. C. Wason. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behaviour*, 4:7–11, 1965.
- [110] P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20:273–281, 1968.

- [111] P.C. Wason. Problem solving. Lemma in R.L. Gregory (ed.), *The Oxford companion to the mind*, Oxford University Press, 1987.
- [112] P. C. Wason and D. W. Green. Reasoning and mental representation. *Quarterly Journal of Experimental Psychology*, 36A:598–611, 1984.
- [113] 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.
- [114] P. C. Wason and P. N. Johnson-Laird. *Psychology of Reasoning: Structure and Content*. Harvard University Press, Boston, 1972.
- [115] P. C. Wason and D. Shapiro. Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23:63–71, 1971.
- [116] A. Whiten (ed.) *Natural theories of mind*. Blackwell, 1991.
- [117] L. Wing. *The Autistic Spectrum: A guide for parents and professionals*. Constable, 1996.
- [118] 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.