

To appear in *Cognitive Science*, 2005.

# A working memory model of relations between interpretation and reasoning\*

Keith Stenning and Michiel van Lambalgen

February 25, 2005

## Abstract

Interpretation is the process whereby a hearer reasons *to* an interpretation of a speaker's discourse. The hearer normally adopts a *credulous* attitude to the discourse, at least for the purposes of interpreting it. That is to say the hearer tries to accommodate the truth of all the speaker's utterances in deriving an intended model. We present a nonmonotonic logical model of this process which defines unique minimal preferred models and efficiently simulates a kind of closed-world reasoning of particular interest for human cognition.

Byrne's 'suppression' data [5] are used to illustrate how variants on this logic can capture and motivate subtly different interpretative stances which different subjects adopt, thus indicating where more fine-grained empirical data are required to understand what subjects are doing in this task.

We then show that this logical competence model can be implemented in spreading activation network models. A one pass process interprets the textual input by constructing a network which then computes minimal preferred models for (3-valued) valuations of the set of propositions of the text. The neural implementation distinguishes easy forward reasoning from more complex backward reasoning in a way that may be useful in explaining directionality in human reasoning.

## 1 Introduction and outline of the paper

This paper proposes a new engagement between logic and the study of the cognitive processes of deductive reasoning. Most concretely we propose a default logical model of some data from Byrne's 'suppression task' [5]. This model reveals how results from the logic programming literature yield the

---

\*Comments from two anonymous referees have helped us to improve the presentation of the paper. The second author is grateful to the Netherlands Organization for Scientific Research (NWO) for support under grant 360-80-000. This paper was inspired by a visit to Waltter Schaeken and Kristien Dieussaert (cf. [11]).

implementability of the model in spreading activation networks. More generally, we show how this logical model changes our construal of the cognitive issues. This proposal connects several seemingly unconnected topics and we owe the reader an outline of their new relations:

(1) We distinguish between two main kinds of logical reasoning: reasoning *from* a fixed interpretation of the logical and nonlogical terms in the premisses, and reasoning *toward* an interpretation of those terms.

(2) We illustrate the distinction by means of Byrne’s ‘suppression task’ [5]. Byrne took her data (a ‘failure’ to apply classical logic) to be evidence against rule-based accounts of logical reasoning and favouring ‘mental models’; but this follows only if experimental subjects in her task reason from a fixed interpretation. If what they are doing is reasoning to a consistent interpretation of the experimental materials, their answers can be shown to make perfect logical sense, albeit in a different logic.

(3) We then show that reasoning *toward* an interpretation has an important feature in common with planning, namely model-construction, and can be treated by the same logical tools, for which we take logic programming with negation as failure. This is a particular form of nonmonotonic logic which has been shown to be extremely useful in discourse interpretation in van Lambalgen and Hamm [46].

(4) We show how the deviation from classical logical reasoning found in Byrne’s data and those of others (e.g. Dieussaert et al. [11]) can be accounted for in the proposed formalism.

(5) We then turn to a possible neural implementation of the proposed formalism for reasoning toward an interpretation (or, alternatively, for planning). Here the semantics of the formalism, Kleene’s strong three-valued logic, becomes prominent, since it allows a slick implementation the logic in terms of isomorphic neural nets coupled by inhibitory connections. We view this implementation as a contribution toward the study of the involvement of the memory systems in reasoning. Also the observed difference in performance between ‘forward’ reasoning patterns (e.g. modus ponens) and ‘backward’ reasoning patterns (e.g. modus tollens) is easily explained in this setup.

(6) The discussion section considers the much-debated issue of the computational complexity of nonmonotonic reasoning, and relates the material in the body of the paper to so-called ‘dual process’ theories of reasoning (for which see e.g. Pollock [28], Stanovich [37] and Evans [13]).

The psychology of reasoning literature alludes to the implication of defeasi-

ble processes in interpretation at a number of points, and there are several different logical and computational frameworks for modelling defeasible reasoning in AI. However, it should be clear that we propose here nothing less than a fundamental change in the standards of the field, and the consequent goals and treatments of evidence. Many psychologists would have us accept that logic is a quite separate endeavour from psychology and that logical models are somehow merely ‘technical’. But it was Johnson-Laird [18] himself who proposed that the psychology of reasoning requires competence models as well as performance models. With this much even the mental logicians can agree. Mental models theory contains within itself various different proposals about competence models for various tasks, but these are rarely clearly separated from the performance modelling, nor, crucially here, even distinguished from each other.

Formal logic is a discipline which studies competence models of reasoning, and which provides a body of systems and mathematical techniques for creating others. The inter-relations between systems are well understood from a century’s worth of study. Psychologists’ failure to exploit this knowledge, the existing systems, and the available techniques for creating new competence models, has led them, with few exceptions, to fail to make basic psychological distinctions (e.g. between defeasible interpretation and monotonic derivation here) and to fail to make simple predictions from the logical distinctions they have made (e.g. that deontic and descriptive conditionals pose quite different problems in the selection task (see Stenning and van Lambalgen[42])). Indeed a substantial proportion of psychological effort has gone into showing that logic cannot provide insight into the basis of human reasoning, with disastrous consequences for the psychology of reasoning. We maintain, on the contrary, that using these logical systems as competence models opens up the possibility of much richer psychological accounts because the variety of systems and their relations allow modelling of the many different things that subjects are doing within and between experiments.

This paper uses formal-logical machinery extensively, and some may think excessively. But for those who feel that the logical technicalities are not needed for a psychological understanding, we offer the following analogy: in the same way as optics is the geometry of light rays, logic is the mathematics of reasoning systems. No one would maintain that the technicalities of optics are irrelevant to how the visual system infers, say, form from motion. Likewise, no consideration of actual human reasoning processes can proceed without careful attention to mathematical constraints on these processes..

## 1.1 Two kinds of reasoning

Our most general aim in this paper is to differentiate interpretation tasks from derivation tasks with which they are sometimes confused, and in doing

so, to establish a productive relation between logical and psychological studies of human reasoning and memory. So we should start by saying what we include under the term interpretation. We intend to use the term broadly for all the structures and processes which connect language to the specifics of the current context.

*All men are mortal. Socrates is a man. So he is mortal.* are two premisses and a conclusion. Without interpretation they have no truth values. Interpreted in the currently relevant range of contexts, both premisses may be true, or one or other or both false e.g. because Socrates is immortal in his appearance in the works of Plato, or where Socrates is a dog. Even if both are true, the ‘he’ in the conclusion may refer to someone other than Socrates. Agreeing about what is going to count as being a man or mortal, who Socrates is in the current argument, and the antecedent of ‘he’ are all parts of the process of interpretation. Under one interpretation anyone who has died is definitely not immortal, whereas under another this may not be the case. Women may or may not be intended to be included in the denotation of ‘men’, and so on.

Our claim in this paper is that central logical concepts such as logical category, truth, falsity and validity can also vary across interpretations; we will therefore refer to these concepts as ‘parameters’. For instance, we propose to apply a logic which makes ‘if ... then’ into something other than a sentential connective, and the notion of validity into ‘truth in all *preferred* models’. If we are right that the logic proposed is an appropriate model for certain kinds of discourse then the parameters characterizing this logic will have to be set in the process of assigning an interpretation, with the psychological consequence that they may also be the subject of misunderstanding when parties make different settings.

Such misunderstandings are, we claim, endemic in the communication between psychologists and their subjects in reasoning tasks. Elsewhere we have argued that Wason’s selection task can be understood as largely invoking interpretative processes in subjects (Stenning and van Lambalgen [41], [42]). Stenning and Cox [39] have shown that syllogistic reasoning has, for the typical undergraduate subject, significant interpretative components, and that these interpretative processes play an important role in determining subsequent reasoning. Here we extend this approach to Byrne’s suppression task[5] – a more obviously interpretative task than either the selection task or syllogisms. The suppression task is a case of what we will call *credulous* text interpretation, where, by definition, listeners engage in credulous text interpretation when they attempt to guess the speaker’s intended model (i.e. a model which makes the speakers’ utterances true and which the speaker intends the hearer to arrive at). A credulous attitude to interpretation contrasts with a sceptical attitude under which a hearer seeks countermodels of the speaker’s statements – models which make all the speaker’s premisses true but their conclusion false.

This contrast between credulous and sceptical attitudes underlies one of the most pervasive logical contrasts between classical monotonic and non-classical nonmonotonic concepts of validity. Thus we wish to put forward a picture of cognition which takes logical categories seriously – one in which participants engaging in discourse are thereby constructing their own interpretations appropriate to their purposes, and this requires the invocation of multiple logical systems.

The history of the confusion between interpretation and derivation processes in psychology is important to understanding the point of the present paper. Logic has always assumed that the process of interpretation of a fragment of language (an argument) into a formal representation is a substantial process, and even traditional logic at least provided a characterisation of what constituted a complete interpretation (and the range of possible such interpretations).

In traditional, pre-symbolic. logic, it was assumed that logic could say little about the process whereby assumptions were adopted, rejected, or modified, or how alternative interpretations of natural languages into artificial ones was achieved. Traditional logical theory narrowly construed was pretty much defined by the limits of what could be said about reasoning *from* interpretations wheresoever those interpretations came from. Nevertheless, interpretation was always assumed to be a substantial part of the process of human reasoning, and much of traditional logical education focussed on learning to detect illicit *shifts* of interpretation within arguments – *equivocations*. It also trained students to distinguish ‘fallacies’ which can be viewed as patterns of assumption easily mistaken for valid patterns of conclusion, thus confusing interpretative with derivational processes.<sup>1</sup> Traditional logic may have had no formal theory of interpretation but it had very prominent place-holders for such an apparatus, and considerable informal understanding of it. It was obvious to all concerned that interpretation was the process whereby content entered into the determination of form. We shall see below, in section 2.1, that interpretation plays a much more explicit role in the modern conception of logic.

## 1.2 The suppression task and its role in the psychology of reasoning

Suppose one presents a subject with the following innocuous premisses:

---

<sup>1</sup>As we shall see, classical derivational fallacies are often valid in nonmonotonic logics of interpretation. This is a cognitively important insight. Rather than having to model fallacies as arbitrarily introduced rules of inference (or reject rules altogether), or for example speculate about ‘real world conditionals often being biconditionals’, seeing fallacies as valid patterns in contrasting logics provides scope for an *explanation* of why a subject with a different construal of the task may have a different notion of validity and so draw different inferences.

- (1) *If she has an essay to write she will study late in the library.*  
*She has an essay to write.* In this case roughly 90% of subjects draw the conclusion ‘She will study late in the library’. Next suppose one adds the premiss
- (2) *If the library is open, she will study late in the library.*

In this case, only 60% concludes ‘She will study late in the library’.

However, if instead of (2) the premiss

- (3) *If she has a textbook to read, she will study late in the library.*

is added, then the percentage of ‘She will study late in the library’ – conclusions is comparable to that in (1).

These observations are originally due to Ruth Byrne [5], and they were used by her to argue against a rule-based account of logical reasoning such as found in, e.g., Rips [32]. For if valid arguments can be suppressed, then surely logical inference cannot be a matter of blindly applying rules; and furthermore the fact that suppression depends on the *content* of the added premiss is taken to be an argument against the role of logical *form* in reasoning. We will question the soundness of the argument, which we believe to rest on two confusions, one concerning the notion of logical form, the other concerning the aforementioned distinction between reasoning from and reasoning to an interpretation. But we agree that the data are both suggestive and important. We will give a brief overview of how these data have been treated in the literature following on from Byrne’s paper, and will provide some reasons to think that these treatments may not be the last word. Readers not interested in the survey can move on to section 1.2.4 without loss.

We start with Byrne’s explanation in [5] of how mental models theory explains the data. ‘Mental models’ assumes that the reasoning process consists of the following stages:

- (i) first the premisses are understood in the sense that a model is constructed on the basis of general knowledge and the specific premisses
- (ii) an informative conclusion is read off from the model
- (iii) this conclusion is checked against possible alternative models of the situation.

In this particular case the model for the premisses  $p \rightarrow q$ ,  $r \rightarrow q$  that is constructed depends on the content of  $p$  and  $r$ , and the general knowledge activated by those contents. If  $r$  represents an alternative, the mental model constructed is one appropriate to the formula  $p \vee r \rightarrow q$ , and the conclusion  $q$  can be read off from this model. If however  $r$  represents an additional condition, that model is appropriate to the formula  $p \wedge r \rightarrow q$ , and no informative conclusion follows.

There are several problems with this explanation and Byrne's use of the suppression effect in the 'rules versus models' debate. The first is that 'rules' and 'models' may not be as mutually exclusive as they are often presented, and in particular there may be *both* 'rules' and 'models' explanations of what is going on [40, 43]. The 'mental rules' account of logical reasoning is falsified only on a somewhat simplistic view of the logical form of the experimental materials. As will be seen below, an account of logical form more in line with current logical theorizing shows that in the paradigm case (2), modus ponens is simply not applicable.

The second problem is that the 'models' explanation seems more a redescription of the data than a theory of the processing that is going on in the subject's mind when she arrives at the model appropriate for  $p \wedge r \rightarrow q$ : exactly *how* does general knowledge lead to the construction of this model? It is a virtue of Byrne's proposed explanation that it generates this question, but we do not think it has been solved by mental modellers. It is precisely the purpose of this paper to propose a solution at several levels, ranging from the logico-symbolic to the neural. We do so because we believe that the suppression task encapsulates important aspects of the process of natural language interpretation, which have wider implications as well (for some of these see van Lambalgen and Hamm [46]; also see the paper [47] on the relation between reasoning patterns and executive defects in autism). In the remainder of this review we discuss some work on the suppression task related to interpretative processes, and we indicate areas where we believe further work (such as that presented here) needs to be done. In order to facilitate the discussion of other contributions it is useful to make some preliminary distinctions.

The benefit of seeking formal accounts of interpretations is that it forces one to be clear about the task as conceived by the subject/experimenter. Unfortunately, few authors have been entirely clear about what they consider to be the (range of) interpretation of the materials which subjects *should* make, let alone what proportions of subjects actually adopt which interpretations. For example, although a considerable number of Byrne's subjects (about 35%) withdraw the modus ponens inference after the second conditional premiss is presented, many more (about 65%) continue to draw the inference. What interpretation of the materials do these subjects have? If it is the same, then why do they not withdraw the inference too? And if it is different, then how can it be accommodated within the semantic framework that underpins the theory of reasoning? Does failure to suppress mean that these subjects have mental logics with inference rules (as Byrne would presumably have interpreted the data if no subject had suppressed)? The psychological data is full of variation, but the psychological conclusions have been rather monolithic.

Lechler [22] used informal interviewing techniques to elicit subjects' likely interpretations of similar materials and showed that a wide range of inter-

pretations is made.

The psychological literature has discussed some varieties of interpretation, and in particular it has distinguished between a *suppositional* and a *probabilistic* interpretation. It is worth our while to discuss these briefly, since the task becomes determinate only when the interpretations are. In the end we shall query how radically these interpretations really differ, but for now the distinction is useful.

### 1.2.1 Suppositional interpretations

Subjects may choose to interpret the task ‘suppositionally’ in the sense of seeking a interpretation of the materials in which all the conditionals are assumed to be absolutely true, and proceeding from there by deduction. Implicitly, Byrne [5] assumes that this is what subjects do, while assuming furthermore that subjects adopt the material implication as their model of the conditional. As we will see below, the combination of the two assumptions has been criticized (rightly, in our view), but it should be pointed out that one may formulate the suppositional interpretation also as: ‘seeking an interpretation of the materials in which all conditionals are assumed to be true *as eventually interpreted*’. The rub is in the italicized phrase: we need not assume that the prima facie interpretation of a discourse is also definitive. In particular the logical form assigned to a statement on first reading (or hearing) may be substituted for another one upon consideration. Only if the proviso expressed by the italicized phrase is disregarded, can the suppression effect be taken to falsify ‘rule’ theories. The reader may be forgiven if at this stage it is unclear how the phrase ‘as eventually interpreted’ can be given formal content; explaining this will be task of the technical part of the paper.

It is of some importance to note that when using the expression ‘deduction’ as above, it is not implied that the mental process involved is derivation in classical logic. What we mean is that the processes are such as to yield an interpretation (or re-interpretation) of the premisses which we can entertain as true (what we call a *credulous* interpretation), from which one can then proceed to deduce consequences by whatever logic one finds appropriate.

There are however circumstances in which we would be well-advised to adopt a non-credulous interpretation. Consider the possibility that the succeeding conditionals are asserted by different speakers than the initial ones, and suppose it is clear that they are intended to voice disagreements:

Speaker A: “If she has an essay, she’s in the library – she’s a very diligent student you know”

Speaker B: “Nonsense! If her boyfriend has called she’ll be at the cinema – she’s an airhead”

Now it would be utterly inappropriate to credulously accomodate. The dialogue must be represented as involving a contradiction and a different,

non-credulous logic is required to do that. Classical logic is useful here, because it allows one to declare a rule to be false if there is a counterexample. By contrast, credulous logics such as the nonmonotonic logic we will be using here do not allow the possibility for falsifying a rule, since putative counterexamples are re-interpreted as exceptions. The following terminology is intended to capture these distinctions. An *absolute-suppositional* interpretation is one in which the discourse-material (here the conditionals) is taken at face-value, without the need or indeed the possibility for re-interpretation. A *suppositional* interpretation *per se* is one in which re-interpretation is allowed.

### 1.2.2 Probabilistic interpretations

The literature on the suppression task contrasts suppositional readings, as explained above, with ‘probabilistic’ readings in which the strength of association between antecedent and consequent of the conditionals is variable. The intuition behind this is that a second conditional premiss may decrease that strength, in which case a subject is no longer willing to draw a conclusion. One way to formalize this idea is to express the conditional ‘if  $A$  then  $B$ ’ as a conditional probability  $P(B | A) = x$ . The task is then interpreted as judging how the arrival of subsequent premisses affects the conditional probabilities of the truth of the candidate conclusions. However, there are other ways of conceiving ‘strength of association’; see for example the discussion of Chan and Chua [7] below.

There is a clear difference, both theoretically and empirically, between *absolute-suppositional* and probabilistic readings; but one outcome of the analysis will be that the probabilistic interpretation can do the work assigned to it only if the subject may re-interpret the materials as postulated in the suppositional interpretation. Armed with these distinctions, we now turn to a brief inventory of some important previous work on the suppression task, in chronological order.

### 1.2.3 Review of work on the suppression task

A number of authors believe that the main problem with the ‘mental models’ account of the suppression effect lies in its assumption of an absolute connection between antecedent and consequent. Thus, Chan and Chua [7] upbraid both the ‘rules’ and ‘models’ camps for failure ‘to give a principled account of the interpretative component involved in reasoning’. Chan and Chua propose a ‘salience’ theory of the suppression task, according to which antecedents are more or less strongly connected with their consequents – ‘operationalised as ratings of perceived importance [of the antecedent in the second conditional premiss] as additional requirements for the occurrence of the consequent [7, p. 222].’ Thus Chan and Chua adopt a ‘weak regular-

ity' interpretation of the conditional instead of the material implication (i.e. more like a conditional probability), and they assume their subjects do too. They correctly argue that both the 'rules' and the 'models' theories do not fit this interpretation, because

[W]e believe that in the suppression experiments, the way subjects cognise the premiss 'If  $R$  then  $Q$ ' may be more fine-grained than merely understanding the antecedent as an additional requirement or a possible alternative.

Their main experimental manipulation accordingly varies the strengths of the connections between antecedents and consequents, and shows that the magnitude of the suppression effect depends upon such variations.

George [15] presents strong evidence that subjects vary as to whether they adopt an absolute-suppositional or a statistical-tendency interpretation of the materials, and that the proportion of interpretations is strongly affected by the response modes offered. By getting subjects to first rate their beliefs in the conditionals used, he showed that about half his subjects adopted absolute-suppositional interpretations in which modus ponens was applied regardless of the subjects' belief in the conditional, whereas for the other half of subjects there was a strong correlation between their degree of belief in, and their willingness to draw conclusions from, conditionals. He further shows that the higher incidence of 'suppression' of the conditional in his population relative to Byrne's is substantially due to offering response categories graded in likelihood. George was the first to observe that instead of being seen as contributing to a debate between mental models and mental logics, Byrne's results point to some 'non-standard forms of reasoning which are outside the scope of both theories'. As far as formal theories of these non-standard patterns goes, George mentions Collins [8] and Collins and Michalski [9].

Stevenson and Over [44] present an account of the suppression effect which is based on a probabilistic interpretation, along with four experiments designed to support the idea that subjects at least *may* interpret the materials of Byrne's experiment in terms of conditional probabilities. Their methods of encouraging a probabilistic interpretation are interesting. Apart from the fact that subjects were asked to make judgements of (qualitative) probabilities of conclusions, in their Experiment 2 the instructions were to imagine that the premisses were part of a conversation *between three people* (our italics). Presumably, having three separate sources was intended to diminish the likelihood of subjects focussing on the sources' intentions about the relations between the statements i.e. to prevent subjects interpreting the statements as a discourse.

While accepting that some subjects do interpret the materials suppositionally, Stevenson and Over believe that this is somehow unnatural and that the probability-based interpretations are more natural. Two quotes give a

flavour of their position. In motivating the naturalness of a probabilistic interpretation they argue that

... it is, in fact, rational by the highest standards to take proper account of the probability of one's premisses in deductive reasoning [44, p. 615].

... performing inferences from statements treated as absolutely certain is uncommon in ordinary reasoning. We are mainly interested in what subjects will infer from statements in ordinary discourse that they may not believe with certainty and may even have serious doubts about [44, p. 621].

In our opinion, at least two issues should be distinguished here – the issue what knowledge and belief plays a role in arriving at an interpretation, and the issue whether the propositions which enter into those interpretations are absolute or probabilistic. We agree that subjects frequently entertain interpretations with propositions which they believe to be less than certain. We agree that subjects naturally and reasonably bring to bear their general knowledge and belief in constructing interpretations. And we agree that basing action, judgement or even belief on reasoning requires consideration of how some discourse relates to the world. But we also believe that subjects can only arrive at less than absolute interpretations (anymore than at absolute interpretations) by some process of constructing an interpretation. Stevenson and Over need there to be a kind of discourse in which statements are already interpreted on some range of models known to both experimenter and subject within which it is meaningful to assign likelihoods. Most basically, hearers need to decide what domain of interpretation speakers intend before they can possibly assign likelihoods.

So how are we to interpret claims that subjects' interpretations are 'probabilistic'? First there is a technical point to be made. Strictly speaking, of course, one is concerned not with the probability of a premiss but with a presumably true statement about conditional probability. In that sense the suppositional and probabilistic interpretation are in the same boat: the authors cannot mean that subjects may entertain serious doubts about their own assignments of conditional probabilities. In fact, in this and other papers on probabilistic approaches to reasoning there is some ambiguity about what is intended. At one point we read about 'probability of one's premisses', only to be reminded later that we should not take this as an espousal of the cognitive reality of probability theory: 'Even relatively simple derivations in that system are surely too technical for ordinary people [44, p. 638].' So there is a real question is to how we are to interpret occurrences of the phrase 'conditional probability'. Incidentally, Oaksford and Chater [25] present models of probabilistic interpretations of the suppression task and show that these models can fit substantial parts of the data from typical suppression experiments. What is particularly relevant in the current context is that these

authors, like Stevenson and Over, do not present the probability calculus as a plausible processing model, but merely as computational-level model in Marr’s sense – that is a model of what the subjects’ mental processes are ‘designed’ to perform. In contrast, the nonmonotonic logic presented here as a computational model for the suppression effect is also intended as a processing model, via the neural implementation given in section 5.

The most important point we want to make about probabilistic interpretations is this: probabilistic and suppositional interpretations share an important characteristic in that they necessarily require the same interpretative mechanisms. For instance, as designers of expert systems well know, it is notoriously difficult to come up with assignments of conditional probabilities which are consistent in the sense that they can be derived from a joint probability distribution. This already shows that the probabilistic interpretation may face the same problems as the suppositional interpretation: both encounter the need to assure consistency.

Now consider what goes into manipulating conditional probabilities of rules, assuming that subjects are indeed able to assign probabilities.<sup>2</sup> Recall that an advantage of the probabilistic interpretation is supposed to be that conditional probabilities can change under the influence of new information, such as additional conditionals. What makes this possible?

Let the variable  $Y$  represent the proposition ‘she studies late in the library’,  $X$  the proposition ‘she has an essay’, and  $Z$ : ‘the library is open’. Probabilistic modus ponens then asks for the probability  $P(Y | X = 1)$ . It is clear that if the first conditional were to be modelled as a conditional probability table of the form  $P(Y | X)$ , and the second as  $P(Y | Z)$ , the setup trivializes, for then the second conditional cannot influence the conditional probability we are interested in.

What presumably happens instead is that the premisses are first integrated into a Bayesian network which represents causal (in)dependencies, and which determines the conditional probabilities that have to be estimated. For instance, the processing of an *additional* premiss as the second conditional is represented by the move from a Bayesian network of the form depicted in figure 1, to one of the form as depicted in figure 2. This structure would make the subject realize that what she can estimate is the conditional probability table for  $P(Y | X, Z)$ , rather than those for  $P(Y | X)$  and  $P(Y | Z)$ . The additional character of the premiss is reflected in the entry in the table which says  $P(Y = 1 | X = 1, Z = 0) = 0$ . As a consequence we have  $P(Y = 1 | X = 1) = P(Y = 1 | X = 1, Y = 1)P(Z = 1)$ , i.e. probabilistic modus ponens will be suppressed in the absence of further information about  $P(Z)$ .

---

<sup>2</sup>We are skeptical, but also think probabilistic approaches to reasoning only make sense under this assumption. In the following we assume subjects are able to construct Bayesian networks as a form of qualitative probabilistic models.



Figure 1: Bayesian network for probabilistic modus ponens

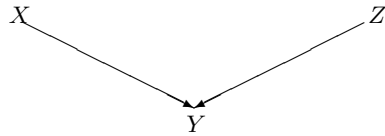


Figure 2: Bayesian network for an additional conditional premiss

Now let us see what happens in the case of an alternative premiss, where  $Z$  is ‘she has a textbook to read’. The specific *alternative* character can only be represented by a different Bayesian network, which now involves an OR variable (defined by the standard truth table), as in figure 3.

More complicated cases can be imagined, involving conditional premisses which are not clearly either additional or alternative. The upshot of this discussion of Stevenson and Over [44] is that moving to a probabilistic interpretation of the premisses does not obviate the need for an interpretative process which (a) constructs a model for the premisses, and (b) conducts a computation on the basis of that model. This is necessary whether the premisses are interpreted probabilistically or logically, although doubtless much more challenging to model in the probabilistic case.

We close with a discussion of a more recent restatement of Byrne’s own position in [6]. Byrne provides a slightly more detailed explanation of the suppression effect in the ‘mental models’ framework, pointing to its emphasis

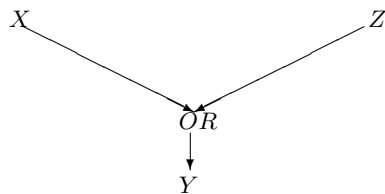


Figure 3: Bayesian network for an alternative conditional premiss

on *counterexamples*

According to the model theory of deduction, people make inferences according to the semantic principle that a conclusion is valid if there are no counterexamples to it [6, p. 350]<sup>3</sup>

and going on to argue against Stevenson and Over that

Our counterexample availability explanation of the suppression effect ... suggests that people disbelieve not the truth of the premisses, but the validity of the conclusion, because they can readily think of a counterexample to it. On our account, people do not doubt the truth of either of the conditional premisses: A conditional that has an antecedent that is insufficient but necessary for the consequent is no more uncertain or doubted than a conditional which has an antecedent that is sufficient but not necessary [6, p. 353].

The explanation thus attempts to stay with a classical concept of validity and, in a roundabout way, to retain classical logic's interpretation of the conditional as material implication as the 'basic' meaning of conditionals. An additional premiss is taken to make subjects aware that the antecedent of the main conditional is not sufficient for its consequent, and subjects are supposed, as a result, to invert the conditional, as evidenced by the non-suppression of the fallacies denial of the antecedent and affirmation of the consequent [6, p. 364]. Indeed, going one step further and adding both an additional and an alternative premiss then suppresses all inferences (Experiment 4, [6, p. 364ff]). The interpretation of this final experiment is interesting:

The suppression of valid inferences and fallacies indicates that people can interpret a conditional within a coherent set of conditionals as supporting none of the standard inferences whatsoever. A conditional can be interpreted as a 'non-conditional,' that is, as containing an antecedent that is neither sufficient nor necessary for its consequent [*ibidem*]

If one now inquires how this 'non-conditional' is defined, one is referred to Table 2 [6, p. 349] where one finds that it is determined by the truth table of a tautology. This we take to be a *reductio ad absurdum* of the attempt to model conditionals using tools from classical logic only, an attempt already made inauspicious by the weight of linguistic evidence against it. As we will see below, nonmonotonic accounts of the conditional allow a more fine-grained study of when and why the four inference patterns are (in)applicable,

---

<sup>3</sup>In principle, a counterexample is different from an exception: the former falsifies a rule, the latter does not. It seems however that Byrne et al. reinterpret counterexamples as exceptions, as the next quotations indicate. This terminological confusion is revealing of the lack of clarity whether mental models is to be interpreted as a classical or as a non-monotonic logic, and of the model-theoretic rhetoric inducing talk of 'examples' where what is involved are generic rules about boundary conditions.

which go much beyond saying that none applies. It therefore comes as something of a surprise to see that the penultimate section of [6] is entitled ‘*Suppression and the non-monotonicity or defeasibility of inferences*’, where it is claimed that ‘the model theory attempts to provide an account of one sort of non-monotonic reasoning, undoing default assumptions ...’ [6, p. 370]. We have been unable to give formal substance to this claim, but if true the ‘mental models’ solution and the one given here should be mutually translatable, and we look forward to seeing such a proof.

#### 1.2.4 Conclusions and outlook

A common thread through the articles reviewed here is that their authors try to explain the suppression effect by assuming a link between the antecedent and consequent of a rule which is different from material implication. This is certainly the correct way to go, but it leaves several questions unanswered. Human reasoning may well mostly be about propositions which are less than certain, and known to be less than certain, but our processes of understanding which meaning is intended have to work initially by credulously interpreting the discourse which describes the situation, and in doing so accommodating apparent inconsistencies, repairs and revisions. For instance, only after we have some specification of the domain can we start consistently estimating probabilities in the light of our beliefs about what we have been asked to suppose. Of course, our judgements about the likely intended interpretation are as uncertain as all our other judgements, but we judge how to accommodate utterances in a way that makes them absolutely consistent.

The good news about the ‘mental models’ approach is that it draws attention to this interpretative process, but it simultaneously fails to distinguish between reasoning *to* and reasoning *from* an interpretation, and it is sorely lacking in detail. The present paper shows that a much more explicit approach to interpretation is possible once one accepts the lessons from the authors discussed above (and from linguistic discussions about the meaning of conditionals; see for example [41, 42, 45, 1]), that subjects generally allow rules to have exceptions. It will be seen that there is in fact an intimate connection between exception-handling and interpretation of discourse. As a first step in that direction, we will now look at an area where efficient exception-handling is of prime importance, namely planning.

### 1.3 Logic, working memory and planning

In some ways classical logic is the nemesis of working memory. Consider what is involved in checking semantically whether an argument of the form  $\varphi_1, \varphi_2 / \psi$  is classically valid. One has to construct a model  $\mathcal{M}$ , then check whether  $\mathcal{M} \models \varphi_1, \varphi_2$ ; if not, discard  $\mathcal{M}$ ; otherwise, proceed to check whether

$\mathcal{M} \models \psi$ , and repeat until all models have been checked. This procedure puts heavy demands on working memory, because the models which have to be constructed are generally not saliently different, so are hard to tell apart. By the same token, it is not easy to check whether one has looked at all relevant models. The fact that classical logic does not fit harmoniously with the operation of working memory already suggests that classical logic is not the most prominent logic for interpreting discourse. Classical logic may indeed be an acquired trick as is sometimes maintained, because it requires overcoming the tyranny of working memory. There may however be other logics which are very much easier on working memory, for instance because the number of models to be considered is much lower, or because these models exhibit salient differences; and we claim that planning provides a good source for such logics.

By definition, planning consists in the construction of a *sequence* of actions which will achieve a given goal, taking into account properties of the world and the agent, and also events that might occur in the world. Both humans and nonhuman primates engage in planning. It has even been attested in monkeys. In recent experiments with squirrel monkeys by McGonigle, Chalmers and Dickinson [24], a monkey has to touch all shapes appearing on a computer screen, where the shapes are reshuffled randomly after each trial. The shapes come in different colours, and the interesting fact is that, after extensive training, the monkey comes up with the plan of touching all shapes of a particular colour, and doing this for each colour. This example clearly shows the hierarchical nature of planning: a goal is to be achieved by means of actions which are themselves composed of actions. It is precisely the hierarchical, ‘recursive’ nature of planning which has led some researchers to surmise that planning has been co-opted by the language faculty, especially syntax (Greenfield [17]; Steedman [38]). There is also a route from planning to language that goes via semantics. There is a live possibility that a distinguishing feature of human language vis à vis ape language is the ability to engage in discourse. Chimpanzees can produce single sentences, which when read charitably show some signs of syntax. But stringing sentences together into a discourse, with all the anaphoric and temporal relations that this entails, seems to be beyond the linguistic capabilities of apes. One can make a good case, however, that constructing a temporal ordering of events out of a discourse involves an appeal to the planning faculty (see van Lambalgen and Hamm [46]).

We defined planning as setting a goal and devising a *sequence* of actions that will achieve that goal, taking into account events in, and properties of the world and the agent. In this definition, ‘will achieve’ cannot mean: ‘*provably* achieves’, because of the notorious frame problem: it is impossible to take into account all eventualities whose occurrence might be relevant to the success of the plan. Therefore the question arises: what makes a good plan? A reasonable suggestion is: the plan works to the best of one’s present

knowledge. Viewed in terms of models, this means that the plan achieves the goal in a ‘minimal model’ of reality, where, very roughly speaking, every proposition is false which you have no reason to assume to be true. In particular, in the minimal model no events occur which are not forced to occur by the data. This makes planning a form of nonmonotonic reasoning: the fact that

‘goal  $G$  can be achieved in circumstances  $C$ ’

does not imply

‘goal  $G$  can be achieved in circumstances  $C + D$ ’

The book van Lambalgen and Hamm [46] formalizes the computations performed by the planning faculty by means of a temporal reasoner (the Event Calculus) as formulated in a particular type of nonmonotonic logic, namely first-order constraint logic programming with negation as failure.<sup>4</sup> Syntactically, logic programming is a good formalism for planning because its derivations are built on backward chaining (regression) from a given goal. Semantically, it corresponds to the intuition that planning consists in part of constructing minimal models of the world. The purpose of [46] is to show that the semantics of tense and aspect in natural language can be formally explained on the assumption that temporal notions are encoded in such a way as to subserve planning. For our present purposes we may abstract from the temporal component of planning, and concentrate on the skeleton of the inference engine required for planning, namely propositional logic programming. Planning proceeds with respect to a model of the world and it is hypothesised that the automatic process of constructing a minimal model which underlies planning also subserves discourse integration. We present our analysis of the suppression task as evidence for this hypothesis.

Nonmonotonic logics abound, of course,<sup>5</sup> but logic programming is attractive because it is both relatively expressive and computationally efficient.<sup>6</sup> Below we shall see that logic programming also has an appealing implementation in neural nets, and that it may thus shed some light on the operation of working memory. Taken together, the proven merits of logic programming in discourse processing [46] and its straightforward implementation in neural networks suggests to us that it is relevant to cognitive modelling. We are not aware of any other nonmonotonic logic which has this

---

<sup>4</sup>This system is originally due to Kowalski and Sergot [20]), with improvements by Shanahan [34, 35, 33, 36], and the authors of [46].

<sup>5</sup>See Pelletier and Elio [26] for a plea for more extensive investigations into the psychological reality of nonmonotonic logics. The present paper proposes that the suppression effect may be used as a testbed.

<sup>6</sup>This is because it does not involve the consistency checks necessary for other nonmonotonic logics, such as Reiter’s default logic [31]. This point is further elaborated in section 6 below.

range of features, but definitely do not claim that there cannot be any such logic.<sup>7,8</sup>

## 2 The suppression effect

The suppression task sets subjects the problem of finding an interpretation which accommodates premises which at least superficially may conflict. It therefore provides a good illustration of how default logic can be used to model human interpretation and reasoning. It is one of the benefits of formalisation that it reveals many aspects of the data which are in need of clarification and suggests how finer-grained data might be found, as well as how the formalism may be modified to account for richer data. As the literature review above revealed, there are several prominent interpretations subjects may adopt, and there are certainly more than was mentioned there. All that is attempted here is to provide an illustrative model of one important interpretation subjects adopt. We believe this is a reasonable reconstruction of how Byrne appears to believe her subjects are interpreting the materials.

As mentioned in the introduction, if one presents a subject with the following premisses:

- (4) a. *If she has an essay to write she will study late in the library.*
- b. *She has an essay to write.*

roughly 90% of subjects<sup>9</sup> draw the conclusion ‘She will study late in the library’ (we will later discuss what the remaining 10% may be thinking). Next suppose one adds the premiss

- (5) *If the library is open, she will study late in the library.*

and one asks again: what follows? In this case, only 60% concludes ‘She will study late in the library’.

However, if instead of the above, the premiss

- (6) *If she has a textbook to read, she will study late in the library*

---

<sup>7</sup>Consider for instance the defeasible planner OSCAR developed by Pollock (see e.g. [29]). This planner is built on top of a theorem prover for classical predicate logic, and is thus more expressive than logic programming. But the gain in expressiveness is paid for by less pleasing computational properties, since such a system cannot be interpreted semantically by iterations of a fixed point operator, a prerequisite for an efficient neural implementation.

<sup>8</sup>An anonymous referee asked whether it might not be possible to model the suppression effect using *assumption-based systems* (see e.g. Poole [30]). The short answer is ‘yes’, because the assumption-based framework is so broad that any nonmonotonic logic, including logic programming, can be modelled in it (see Bondarenko et al. [4]). But the translation does not necessarily increase perspicuity.

<sup>9</sup>The figures we use come from the experiment reported in Dieussaert et al. [11].

is added, then the percentage of ‘She will study late in the library’-conclusions is around 95%.

In this type of experiment one investigates not only *modus ponens* (MP), but also *modus tollens* (MT), and the ‘fallacies’ *affirmation of the consequent* (AC), and *denial of the antecedent* (DA), with respect to both types of added premisses, (5) and (6). In Table 1 we tabulate the relevant data, following Dieussaert et al. [11], since the experiments reported in this study have more statistical power than those of Byrne [5].

## 2.1 Logical form

The conclusion that Byrne draws from the experimental results is that

... in order to explain how people reason, we need to explain how premisses 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 [5, p. 83].

Byrne thus sees the main problem as explaining ‘how premisses of the same apparent logical form can be interpreted in quite different ways’. We would question the accuracy of this formulation, and instead prefer to formulate the main issue as follows: it is the job of the interpretative process to *assign* logical form, which cannot simply be read off from the given material. In other words, the difference between Byrne and ourselves appears to be this: whereas she takes logical form to be more or less given, we view it as the end result of a (possibly laborious) interpretative process. This difference is connected to a different view of what logical form is. Psychologists have often taken this to mean the formal expression which results from translating the surface structure of the given sentence into a chosen formal language; this is apparently how Byrne conceives of logical form. However, from a logician’s point of view much more is involved. Let  $\mathcal{N}$  be (a fragment of) natural language. A more complete list<sup>10</sup> of what is involved in assigning logical form to expressions in  $\mathcal{N}$  is given by:

1.  $\mathcal{L}$  a formal language into which  $\mathcal{N}$  is translated
2. the expression in  $\mathcal{L}$  which translates an expression in  $\mathcal{N}$
3. the semantics  $\mathcal{S}$  for  $\mathcal{L}$
4. the definition of validity of arguments  $\psi_1, \dots, \psi_n/\varphi$ , with premisses  $\psi_i$  and conclusion  $\varphi$ .

We can see from this list that assigning logical form is a matter of setting parameters. For each item on the list, there are many possibilities for variation. Take just one example, the choice of a formal language. One possibility here is the ordinary recursive definition, which has clauses like ‘if  $A, B$  are formulas, then so is  $A \rightarrow B$ ’, thus allowing for iteration of the conditional. However, another possibility, and one which we shall choose, is where formation of  $A \rightarrow B$  is restricted to  $A, B$  which do not themselves contain a conditional. Furthermore, these parameters are independent: if one has decided to translate  $\mathcal{N}$  into a propositional language with connectives  $\neg, \vee, \wedge, \rightarrow$ , one is still at liberty to choose a semantics for these connectives;

---

<sup>10</sup>See [42, section 1.1] for discussion.

and, perhaps more surprisingly, one is also at liberty to choose a definition of validity. The classical definition of validity: ‘an argument is valid if the conclusion is true in all models of the premisses’, is but one possibility; the general form of a nonmonotonic notion of validity: ‘an argument is valid if the conclusion is valid in all *preferred* models of the premisses’ is another. In fact, applying the classical definition of validity means that one must leave out of consideration all information we happen to have, beyond the premisses. This is typically almost impossible to achieve for those without logical training.<sup>11</sup>

In the remainder of the paper we will reanalyse the data regarding the suppression effect with the above distinctions in mind, and show their relevance for the process of interpretation and its relation to working memory. One outcome of the analysis will be that there is no ‘suppression effect’ in the sense of suppression of formal reasoning – indeed the observed reasoning patterns, ‘fallacies’ included, conform to well-known logical forms. We will illustrate the main logical ideas involved using (propositional) logic programming with negation as failure, to which we give a brief introduction in section 3.<sup>12</sup> Before we do so, we briefly discuss some aspects of the meaning of the natural language conditional.

## 2.2 A logical form for the conditional

We have seen above that several authors have tried to explain the suppression effect by assuming that the conditional expresses a regularity rather than a material implication, as in classical logic. Furthermore, the literature on the semantics of natural language provides a wealth of data showing that the identification of the conditional with the material implication is not generally warranted. It is impossible to do justice to that vast literature here, so we content ourselves with references to two books containing important review articles: [45] and [1].

The meaning of the conditional that we shall focus on is that of a law-like relationship between antecedent and consequent. An example of a law expressed by means of a conditional is

(7) If a glass is dropped on a hard surface, it will break.

or

(8) If a body is dropped, its velocity will increase as  $gt^2$ .

What is common to both examples is that the antecedent hides an endless number of unstated assumptions: in the first case, e.g. that the glass is not caught before it falls, etc., in the second case, e.g. that there are no other

---

<sup>11</sup>Byrne replaced all subjects in her sample who had taken a course in logic.

<sup>12</sup>For a fuller introduction, consult [12].

forces at work on the body, etc..<sup>13</sup> We will therefore give the general logical form of lawlike conditionals ‘if  $A$  then  $B$ ’ as

(9) If  $A$ , and nothing abnormal is the case, then  $B$ .

where what is abnormal is provided by the context; we will shortly see examples of this in the suppression task. The preceding formulation, however, still explains ‘if ... then’ in terms of ‘if ... then’, so we must now inquire seriously into the meaning of the conditional. We contend that the conditional is often not so much a truth functional connective, as a license for certain inferences.<sup>14</sup> One reason is the role of putative counterexamples, i.e. situations where  $A$  and not- $B$ ; especially in the case of lawlike conditionals, such a counterexample is not used to discard the conditional, but to look for an abnormality; it is thus more appropriate to describe it as an exception. Thus, one use of the conditional is where it is taken as given, not as a statement which can be false, and we claim that this is the proper way to view the conditionals occurring in the suppression task, which are after all supplied by the experimenter.<sup>15</sup> Stenning and van Lambalgen [41, 42] use these same observations about natural language conditionals to explain many apparently unrelated phenomena in Wason’s selection task.

Having posited that, in the present context, the conditional is rather a license for inferences than a connective, we must determine what these inferences are. One inference is, of course, *modus ponens*: the premisses  $A$  and ‘if  $A$  then  $B$ ’ license the inference that  $B$ . The second type of inference licensed by the conditional may be dubbed ‘closed world reasoning’: it says that if it is impossible to derive a proposition  $B$  from the given premisses

---

<sup>13</sup>In this respect, the conditional provides an interesting contrast to the universal quantifier, with which it is often aligned. To slightly adapt an example due to Nelson Goodman: one can say

(i) All the coins in my pocket are copper.

in order to express a contingent generalisation, but one would not so readily, with the same intent, say

(ii) If a coin is in my pocket, it is copper.

precisely because it is hard to imagine the law-like connection between antecedent and consequent which the conditional suggests. The attempt to interpret this tends to conjure scenarios of reverse alchemy, in which gold coins just moved into the pocket turn to copper.

<sup>14</sup>A connective differs from a license for inference in that a connective, in addition to licensing inferences, also comes with rules for inferring a formula containing that connective as main logical operator.

<sup>15</sup>The preceding considerations imply that the conditional cannot be iterated. Natural language conditionals are notoriously hard (although not impossible) to iterate, especially when a conditional occurs in the antecedent of another conditional – ‘If, if conditionals are iterated, then they aren’t meaningful, then they aren’t material’ is an example; one more reason why ‘if ... then’ is not simply material implication.

by repeated application of *modus ponens*, then one may assume  $B$  is false. This kind of reasoning is routinely applied in daily life: if the timetable says there is no train leaving Edinburgh for London between 5 pm and 5.10 pm, one assumes there is no such train scheduled. Closed world reasoning is what allows humans to circumvent the notorious frame problem (at least to some extent): my reaching for a glass of water may be unsuccessful for any number of reasons, for instance because the earth's gravitational field changes suddenly, but since I have no positive information that this will happen, I assume it will not. We hypothesise that closed world reasoning plays an important part in reasoning with conditionals; in particular, the suppression effect will be seen to be due to a special form of a closed world assumption. We now recast the preceding considerations as a formal definition in logic programming. The next section is inevitably somewhat technical.

### 3 Logic programming

A good framework for a conditional with the properties outlined above is *logic programming*, a fragment of propositional (or predicate) logic with a special, nonclassical, semantics. For the sake of clarity we start with a fragment of logic programming in which negation is not allowed; but note that the proposed notion of conditionals does require negation.

**Definition 1** *A (positive) clause is a formula of the form  $p_1, \dots, p_n \rightarrow q$ , where the  $q, p_i$  are propositional variables; the antecedent may be empty. In this formula,  $q$  is called the head, and  $p_1, \dots, p_n$  the body of the clause. A (positive) program is a finite set of positive clauses.*

Until further notice, we assume that propositions are either true (1) or false (0), but the semantics is nonetheless nonclassical. The only models to be considered are those of the following form

**Definition 2** *Let  $P$  be a positive program on a finite set  $L$  of proposition letters. An assignment  $\mathcal{M}$  of truthvalues  $\{0, 1\}$  to  $L$  (i.e. a function  $\mathcal{M} : L \rightarrow \{0, 1\}$ ) is a model of  $P$  if for  $q \in L$ ,*

1.  $\mathcal{M}(q) = 1$  if there is a clause  $p_1, \dots, p_n \rightarrow q$  in  $P$  such that for all  $i$ ,  $\mathcal{M}(p_i) = 1$
2.  $\mathcal{M}(q) = 0$  if for all clauses  $p_1, \dots, p_n \rightarrow q$  in  $P$  there is some  $p_i$  for which  $\mathcal{M}(p_i) = 0$ .

The definition entails that for any  $q$  not occurring as the head of a clause,  $\mathcal{M}(q) = 0$ . More generally, the model  $\mathcal{M}$  is minimal in the sense that a proposition not forced to be true by the program is false in  $\mathcal{M}$ ; this is our first (though not final) formulation of the closed world assumption.

We will next liberalize the preceding definitions and allow negation in the body of a clause.

**Definition 3** 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>16</sup>, and  $q$  is a propositional variable. Facts are clauses of the form  $\top \rightarrow q$ , which will usually be abbreviated to  $q$ . Empty antecedents are no longer allowed. A definite logic program is a finite conjunction of definite clauses.

In order to give a semantics for negation, the closed world assumption is internalised to what is known as ‘negation as failure’:  $\neg\varphi$  is true if the attempt to derive  $\varphi$  from the program  $P$  fails. The proper semantics for definite programs requires a move from two-valued logic to Kleene’s strong three-valued logic (introduced in [19, p. 332ff]), which has the truth values *undecided* ( $u$ ), *false* (0) and *true* (1). The meaning of *undecided* is that the truth value can evolve toward either true or false (but not conversely).<sup>17</sup> This semantics will be introduced in greater detail below. It is an important fact, however, that the models of interest can be captured by means of the following construction

**Definition 4** (a) The completion of a program  $P$  is given by the following procedure:

1. take all clauses  $\varphi_i \rightarrow q$  whose head is  $q$  and form the expression  $\bigvee_i \varphi_i \rightarrow q$ <sup>18</sup>
2. replace the  $\rightarrow$ ’s by  $\leftrightarrow$ ’s (here,  $\leftrightarrow$  has a classical interpretation given by:  $\psi \leftrightarrow \varphi$  is true if  $\psi, \varphi$  have the same truth value, and false otherwise).
3. this gives the completion of  $P$ , which will be denoted by  $\text{comp}(P)$ .

(b) If  $P$  is a logic program, define the nonmonotonic consequence relation  $\approx$  by

$$P \approx \varphi \text{ iff } \text{comp}(P) \models \varphi.$$

If  $P \approx \varphi$ , we say that  $\varphi$  follows from  $P$  by negation as failure, or by closed world reasoning. The process of completion is also referred to as minimisation.<sup>19</sup>

<sup>16</sup>We use  $\top$  for an arbitrary tautology, and  $\perp$  for an arbitrary contradiction

<sup>17</sup>This is therefore different from saying that the truth value is fuzzy, which requires a linear order of truth values, with *between* 0 and 1.

<sup>18</sup>In the customary definitions it is assumed that if there is no such  $\varphi_i$ , then the expression  $\perp \rightarrow q$  is added. This case cannot occur here.

<sup>19</sup>Readers who recall McCarthy’s use of an abnormality predicate  $ab$  may wonder why we do not use circumscription [23] instead of logic programming, since circumscription also proceeds by minimising the extension of  $ab$ . There is a technical reason for this: circumscription cannot be easily used to explain the fallacies; but the main reason is that logic programming unlike circumscription allows incremental computation of minimal models, and this computation will be seen to be related to the convergence toward a stable state of a neural network. This shows why model-construction can here proceed automatically.

Using the terminology introduced above, our main hypothesis in explaining Byrne’s data as conforming to the logical competence model of closed world reasoning can then be stated as

- (a) the conditionals used in the suppression task, far from being material implications, can be captured much more adequately by logic programming clauses of the form  $p \wedge \neg ab \rightarrow q$ , where  $ab$  is a proposition letter indicating that something abnormal is the case;
- (b) when making interpretations, subjects usually do not consider *all* models of the premisses, but only *minimal* models, defined by a suitable completion.

The readers who wish to see some applications of these notions to the suppression task before delving into further technicalities, may now jump ahead to section 4, in particular the explanations with regard to the forward inferences MP and DA. The backward inferences MT and AC require a slightly different application of logic programming, which will be introduced in the next subsection. The final subsection will look at the construction of models, which is necessary for the connection with neural networks.

### 3.1 A strengthening of the closed world assumption: integrity constraints

So far we have applied the closed world assumption to atomic formulas (for instance  $ab$ ) and their negations: if  $ab$  is not in the database, we may assume it is false. We now extend the closed world assumption to cover program clauses as well: if  $\varphi_1 \rightarrow q, \dots, \varphi_n \rightarrow q$  are all the clauses with nontrivial body which have  $q$  as head, then we (defeasibly) conclude that  $q$  can only be the case *because* one of  $\varphi_1, \dots, \varphi_n$  is the case. We therefore do not consider the possibility that  $q$  is the case because some other state of affairs  $\psi$  obtains, where  $\psi$  is independent of the  $\varphi_1, \dots, \varphi_n$  and such that  $\psi \rightarrow q$ . This kind of reasoning might be called ‘diagnostic’.

This notion of closed world is not quite expressed by the completion of a program. For in the simple case where we have only a clause  $p \rightarrow q$  and a fact  $q$  is added, the completion becomes  $(p \vee \top) \leftrightarrow q$ , from which nothing can be derived about  $p$ . What we need instead is a principled way of adding  $q$  such that the database or model is updated with  $p$ . The proper technical way of achieving this is by means of so-called *integrity constraints*. To clarify this notion, we need a small excursion into database theory, taking an example from Kowalski [21, p. 232].

An *integrity constraint* in a database expresses obligations and prohibitions that the states of the database must satisfy if they fulfill a certain condition. For instance, the ‘obligation’ to carry an umbrella when it is raining may be formalized (using a self-explanatory language for talking about

actions and their effects) by the integrity constraint

$$\text{Holds}(\text{rain}, t) \rightarrow \text{Holds}(\text{carry} - \text{umbrella}, t). \quad (1)$$

The crucial point here is the meaning of  $\rightarrow$ . The formula 1 cannot be an ordinary program clause, for in that case the addition of  $\text{Holds}(\text{rain}, t)$  would trigger the consequence  $\text{Holds}(\text{carry} - \text{umbrella}, t)$  which may well be false, and in any case does not express an obligation.

A better way to think of an integrity constraint is to view the consequent as a constraint that the database must satisfy if the antecedent holds. This entails in general that the database has to be *updated* with a true statement about the world. First a piece of terminology. A formula  $\varphi$  is used as a *query*, and denoted  $?\varphi$ , if one tries to determine whether  $\varphi$  follows from a program  $P$ . The ‘success’ or ‘failure’ of a query is usually defined syntactically, but we will not introduce a derivational apparatus here, and provide only a semantic characterization: a query  $?\varphi$  succeeds with respect to a program  $P$ , if  $\text{comp}(P) \models \varphi$ , i.e. if  $\varphi$  is entailed by the completion of  $P$ . Likewise, a query  $?\varphi$  fails with respect to a program  $P$ , if  $\text{comp}(P) \models \neg\varphi$ .

To return to our example, there will be an action *take* – *umbrella*, linked to the rest of the database by

$$\text{Initiates}(\text{take} - \text{umbrella}, \text{carry} - \text{umbrella}, t).$$

Suppose the database contains  $\text{Holds}(\text{rain}, \text{now})$ , then the integrity constraint requires us to update the database in such a way that the query

$$?\text{Holds}(\text{carry} - \text{umbrella}, \text{now}),$$

succeeds. The appropriate way to do so is, of course, to take an umbrella and inform the database that one has done so.

It is also possible to have an integrity constraint without a condition, such as  $\text{Holds}(\text{rain}, t)$  in the above example. An entry in someone’s diary like ‘appointment in Utrecht, Friday at 9.00’ expresses an unconditional obligation to satisfy  $\text{HoldsAt}(\text{be-in-Utrecht}, \text{Friday-at-9.00})$ , and presented with this integrity constraint, the internal database comes up with a plan to satisfy the constraint. Such unconditional integrity constraints are especially useful for the kind of examples discussed here. Readers who wish to see applications to the suppression task may now jump ahead to the second part of section 4, which treats the backward inferences MT and AC. The next (and final) subsection is only essential for the neural implementation.

### 3.2 Constructing models

In this last subsection, we explain how minimal models can be efficiently computed given a definite logic program. As above we start with a simpler case. Recall that a positive logic program has clauses of the form  $p_1 \wedge \dots \wedge$

$p_n \rightarrow q$ , where the  $p_i, q$  are proposition letters and the antecedent (also called the body of the clause) may be empty. Models of a positive logic program  $P$  are given by the fixed points of a monotone<sup>20</sup> operator:

**Definition 5** *The operator  $T_P$  associated to  $P$  transforms a model  $\mathcal{M}$  (viewed as a function  $\mathcal{M} : L \rightarrow \{0, 1\}$ , where  $L$  is the set of proposition letters) into a model  $T_P(\mathcal{M})$  according to the following stipulations: if  $v$  is a proposition letter,*

1.  $T_P(\mathcal{M})(v) = 1$  if there exists a set of proposition letters  $C$ , true on  $\mathcal{M}$ , such that  $\bigwedge C \rightarrow v \in P$
2.  $T_P(\mathcal{M})(v) = 0$  otherwise.

**Definition 6** *An ordering  $\subseteq$  on models is given by:  $\mathcal{M} \subseteq \mathcal{N}$  if all proposition letters true in  $\mathcal{M}$  are true in  $\mathcal{N}$ .*

**Lemma 1** *If  $P$  is a positive logic program,  $T_P$  is monotone in the sense that  $\mathcal{M} \subseteq \mathcal{N}$  implies  $T_P(\mathcal{M}) \subseteq T_P(\mathcal{N})$ .*

This form of monotonicity would fail if a body of a clause in  $P$  contains a negated atom  $\neg q$  and also a clause  $\neg q \rightarrow s$ : one can then set up things in such a way that  $s$  is true at first, and becomes false later. Hence we will have to complicate matters a bit when considering negation, but this simple case illustrates the use of monotone operators. Monotonicity is important because it implies the existence of so called *fixed points* of the operator  $T_P$ .

**Definition 7** *A fixed point of  $T_P$  is a model  $\mathcal{M}$  such that  $T_P(\mathcal{M}) = \mathcal{M}$ .*

**Lemma 2** *If  $T_P$  is monotone, it has a least and a greatest fixed point. The least fixed point will also be called the minimal model.*

Monotonicity is also important because it allows incremental computation of the minimal model. As noted above, by itself circumscription does not give this computational information. Computability is a consequence of the syntactic restrictions on logic programs.

So far we have only modelled positive programs, but the logic programs that we need must allow negation in the body of a clause, since we model the conditional ‘ $p$  implies  $q$ ’ by the clause  $p \wedge \neg ab \rightarrow q$ . As observed above, extending the definition of the operator  $T_P$  with the classical definition of negation would destroy its monotonicity, necessary for the incremental approach to the least fixed point. One solution is to replace the classical two-valued logic by a particular form of three-valued logic, Kleene’s strong three-valued logic, designed for modelling the process whereby reasoning algorithms take us from uncertainty to definitive values [19, p. 332ff]. This

---

<sup>20</sup>Monotonicity in this sense is also called continuity.

		$p$	$q$	$p \wedge q$	$p$	$q$	$p \vee q$
		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	$u$

Figure 4: Three-valued connectives

logic has truth values  $\{u, 0, 1\}$  with the partial order  $u \leq 0$  and  $u \leq 1$ . Here,  $u$  is *not* a degree of truth, but rather means that the truth value is, so far, undecided. The chosen ordering reflects the intuition that  $u$  can ‘evolve’ toward 0 or 1 as a result of computation. The truth tables given in figure 4 (taken from [19, p. 334]) are then immediate, as readers should satisfy themselves.

In addition we define an equivalence  $\leftrightarrow$  by assigning 1 to  $\varphi \leftrightarrow \psi$  if  $\varphi, \psi$  have the same truth value (in  $\{u, 0, 1\}$ ), and 0 otherwise.

We show how to construct models for such programs, as fixed points of a three-valued consequence operator  $\mathcal{T}_P^3$ . We will drop the superscript when there is no danger of confusion with its two-valued relative defined above.

**Definition 8** *A three-valued model is an assignment of the truth values  $u, 0, 1$  to the set of proposition letters. If the assignment does not use the value  $u$ , the model is called two-valued. If  $\mathcal{M}, \mathcal{N}$  are models, the relation  $\mathcal{M} \leq \mathcal{N}$  means that the truth value of a proposition letter  $p$  in  $\mathcal{M}$  is less than or equal to the truth value of  $p$  in  $\mathcal{N}$  in the canonical ordering on  $u, 0, 1$ .*

**Definition 9** *Let  $P$  be a program.*

- a. The operator  $\mathcal{T}_P$  applied to formulas constructed using only  $\neg, \wedge$  and  $\vee$  is determined by the above truth tables.*
- b. Given a three-valued model  $\mathcal{M}$ ,  $T_P(\mathcal{M})$  is the model determined by*
  - (a)  $T_P(\mathcal{M})(q) = 1$  iff there is a clause  $\varphi \rightarrow q$  such that  $\mathcal{M} \models \varphi$*
  - (b)  $T_P(\mathcal{M})(q) = 0$  iff there is a clause  $\varphi \rightarrow q$  in  $P$  and for all such clauses,  $\mathcal{M} \models \neg\varphi$*

The preceding definition ensures that unrestricted negation as failure applies only to proposition letters  $q$  which occur in a formula  $\perp \rightarrow q$ ; other

proposition letters about which there is no information at all may remain undecided.<sup>21</sup> This will be useful later, when we will sometimes want to restrict negations as failure to  $ab$ . Once a literal has been assigned value 0 or 1 by  $T_P$ , it retains that value at all stages of the construction; if it has been assigned value  $u$ , that value may mutate into 0 or 1 at a later stage.

**Lemma 3** *If  $P$  is a definite logic program,  $T_P$  is monotone in the sense that  $\mathcal{M} \leq \mathcal{N}$  implies  $T_P(\mathcal{M}) \leq T_P(\mathcal{N})$ .*

Here are three essential results, which will turn out to be responsible for the efficient implementability in neural networks.

**Lemma 4** *Let  $P$  be a program.*

1.  $\mathcal{M}$  is a model of the  $\text{comp}(P)$  iff it is a fixed point of  $T_P$ .
2. The least fixed point of  $T_P$  is reached in finitely many steps ( $n + 1$  if the program consists of  $n$  clauses).

**Lemma 5** *If  $P$  is a definite logic program,  $T_P$  is monotone in the sense that  $\mathcal{M} \leq \mathcal{N}$  implies  $T_P(\mathcal{M}) \leq T_P(\mathcal{N})$ .*

**Lemma 6** 1. *The operator  $T_P^3$  has a least fixed point. The least fixed point of  $T_P^3$  will be called the minimal model of  $P$ .*

2. *All models  $\mathcal{M}$  of  $\text{comp}(P)$  are fixed points of  $T_P^3$ , and every fixed point is a model.*

In this context, the nonmonotonic consequence relation  $P \approx \varphi$  (see definition 4) is given by ‘ $\text{comp}(P) \models_3 \varphi$ ’, or in words: all (three-valued) models of  $\text{comp}(P)$  satisfy  $\varphi$ . Observe that the relation  $\approx$  is completely determined by what happens on the least fixed point. Larger fixed points differ in that some values  $u$  in the least fixed point have been changed to 0 or 1 in the larger fixed point; but by the monotonicity property (with respect to truth values) of Kleene’s logic this has no effect on the output unit pairs, in the sense that an output value 1 cannot be changed into 0 (or conversely).

## 4 How this explains nonclassical answers in the suppression task

The explanation of Byrne’s data will be presented in two stages, corresponding to the forward inferences (MP and DA) and the backward inferences (MT and AC).

---

<sup>21</sup>This parallels a similar proviso in the definition of the completion.

Before we present the explanation we want to make a remark of a methodological nature. Once one acknowledges that there are many alternative logical models besides classical logic, logic takes on a more interesting combination of normative and descriptive roles in empirical investigations. So, *if* subjects adopt the processing goals implicit in our proposed nonmonotonic logic, then ‘correct’ performance must be judged by that model. This offers us the possibility of explaining, for example, two groups of subjects drawing different conclusions in one of Byrne’s conditions as *both* conforming to competence models – but different ones. Obviously, if there are sufficient unconstrained competence models to fit any data, then our enterprise is not an empirical one. But the reverse is true. We can readily seek corroborating data by seeing whether subjects’ discourse-processing goals (their construal of the task) accords with the strict assumptions of the proposed competence models. In particular it seems that subjects can have two main discourse-processing goals in this type of task: accomodating the speaker’s utterances, or questioning them. As we have seen, classical logic is appropriate to the latter goal, but not to the former. Below, we will only explain what the appropriate pattern of answers is on the assumption that subjects are accomodating. The reader will readily supply the appropriate answers for the other alternative<sup>22</sup>.

#### 4.1 The forward inferences: MP and DA

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.

**Definition 10** *For our purposes, a program is a finite set of conditionals of the form  $A_1 \wedge \dots \wedge A_n \wedge \neg ab \rightarrow B$ , together with the clauses  $\perp \rightarrow ab$  for all proposition letters of the form  $ab$  occurring in the conditionals. Here, the  $A_i$  are propositional variables or negations thereof, and  $B$  is a propositional variable. We also allow the  $A_i$  to be  $\top$  and  $\perp$ . Empty antecedents are not allowed.*<sup>23</sup>.

In the following we will therefore represent (10-a) as (10-b)

- (10)    a.    If she has an essay, she will study late in the library.  
           b.     $p \wedge \neg ab \rightarrow q$

and (11-a) and (11-b) both as (11-c)

---

<sup>22</sup>We do not have a model which explains the exact distribution of frequencies of answers. If the present analysis is correct, such a model would require variables determining discourse-processing goals.

<sup>23</sup>As mentioned before, this definition is formulated because we do not necessarily want to apply closed world reasoning to *all* proposition letters, although always to proposition letters of the form  $ab$ .

- (11) a. If the library is open, she will study late in the library.  
 b. If she has a textbook to read, she will study late in the library.  
 c.  $r \wedge \neg ab' \rightarrow q$

It is essential that the conditionals are represented as being part of a definite logic program, so that they function as licenses for inference rather than truthfunctional connectives. We show that on the basis of this interpretation, the forward inferences MP and DA *and* their ‘suppression’ correspond to valid argument patterns. The main tool used here is the completion of a program, as a formalisation of the closed world assumption as applied to facts. We emphasise again that in virtue of lemmas 4 – 6, the completion is really shorthand for a particular model. Thus, *what we will be modelling is how subjects reason toward an interpretation of the premisses by suitably adjusting the meaning of the abnormalities*. Once they have reached an interpretation, reasoning from that interpretation is trivial.

The ‘backward’ inferences (MT and AC) require the closed world assumption as applied to rules, and will be treated separately, in section 4.2.

**MP for a single conditional premiss** Suppose we are given a single conditional  $p \wedge \neg ab \rightarrow q$  and the further information that  $p$  (i.e.  $\top \rightarrow p$ ). The full logic program for this situation is  $\{p; p \wedge \neg ab \rightarrow q; \perp \rightarrow ab\}$ . Closed world reasoning as formalised in the completion gives the set  $\{p; p \wedge \neg ab \leftrightarrow q; \perp \leftrightarrow ab\}$ , which is equivalent to  $\{p; p \leftrightarrow q\}$ , from which  $q$  follows. This argument can be rephrased in terms of our distinction between two forms of reasoning as follows.

Reasoning to an interpretation starts with the general form of a conditional<sup>24</sup>, and the decision to apply nonmonotonic closed world reasoning. As a consequence, one derives as the logical form of the conditional  $p \leftrightarrow q$ . Reasoning from an interpretation then starts from this logical form and the atomic premiss  $p$ , and derives  $q$ .

**A ‘fallacy’: DA for a single conditional premiss** Suppose we are again given a conditional  $p \wedge \neg ab \rightarrow q$  and the further information  $\neg p$ .<sup>25</sup> Reasoning to an interpretation starts from the program  $\{\neg p; p \wedge \neg ab \rightarrow q; \perp \rightarrow ab\}$  and the closed world assumption. As above, the end result of that reasoning process is  $\{\neg p; p \leftrightarrow q\}$ . Reasoning from an interpretation then easily derives  $\neg q$ .

<sup>24</sup>There is nothing sacrosanct about this starting point. It may itself be the consequence of a reasoning process, or a different starting point, i.e. a different interpretation of the conditional may be chosen. The formalisation chosen is a first approximation to the idea that natural language conditionals allow exceptions. For other purposes more complex formalisations may be necessary.

<sup>25</sup>Strictly speaking  $\neg p$  is not an allowed clause, but we may interpret  $\neg p$  as obtained from the allowed clause  $\perp \rightarrow p$  by closed world reasoning.

**MP in the presence of an additional premiss** 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  $p \leftrightarrow q$ . Now suppose that the possibility of an abnormality is made salient, e.g. by adding a premiss ‘if the library is open, she will study late in the library’ in Byrne’s example. We propose that this results in the addition of a clause such as  $\neg r \rightarrow ab$ , because the possibility of an  $ab$ -normality is highlighted by the additional premiss.<sup>26</sup> Although this is not essential, the situation may furthermore be taken to be symmetric, in that the first conditional highlights a possible abnormality relating to the second conditional. The circumstance that the library is open is not a sufficient incentive to go and study there, one must have a purpose for doing so.<sup>27</sup> This means that the further condition  $\neg p \rightarrow ab'$  for  $ab'$  is added. That is, reasoning toward an interpretation starts with the set

$$\{p; p \wedge \neg ab \rightarrow q; r \wedge \neg ab' \rightarrow q; \perp \rightarrow ab; \perp \rightarrow ab'; \neg r \rightarrow ab; \neg p \rightarrow ab'\}.$$

Applying closed world reasoning in the form of the completion yields

$$\{p; (p \wedge \neg ab) \vee (r \wedge \neg ab') \leftrightarrow q; (\perp \vee \neg r) \leftrightarrow ab; (\perp \vee \neg p) \leftrightarrow ab'\},$$

which reduces to  $\{p; (p \wedge r) \leftrightarrow q\}$ . Reasoning from an interpretation is now stuck in the absence of information about  $r$ .

Here we see nonmonotonicity at work: the minimal model for the case of an additional premiss is essentially different from the minimal model of a single conditional premiss plus factual information.

**DA in the presence of an additional premiss** Now suppose we have as minor premiss  $\neg p$  instead of  $p$ . Reasoning toward to an interpretation derives as above the set  $\{\neg p; (p \wedge r) \leftrightarrow q\}$ . Since  $\neg(p \wedge r)$ , reasoning from an interpretation concludes  $\neg q$ . It follows that ‘denying the antecedent’ will not be suppressed, as observed.

**MP and DA for an alternative premiss** The difference between this case and the previous one is that, by general knowledge, the alternatives do not highlight possible obstacles. This means that the clauses  $\neg r \rightarrow ab; \neg p \rightarrow ab'$  are lacking. Reasoning to an interpretation thus starts with the set

$$\{p; p \wedge \neg ab \rightarrow q; r \wedge \neg ab' \rightarrow q; \perp \rightarrow ab; \perp \rightarrow ab'\}$$

Closed world reasoning converts this to

$$\{p; (p \wedge \neg ab) \vee r \wedge \neg ab' \leftrightarrow q; \perp \leftrightarrow ab; \perp \leftrightarrow ab'\},$$

---

<sup>26</sup>Obviously, this is one place where general knowledge of content enters into the selection of appropriate logical form. Nothing in the *form* of the sentence tells us that the library being open is a boundary condition on her studying late in the library.

<sup>27</sup>Evidence for such symmetry can be found in Experiment 1 of Byrne et al. [6].

which reduces to  $\{p; (p \vee r) \leftrightarrow q\}$ . Reasoning from an interpretation then easily derives  $q$ : no suppression.

Now consider ‘denial of the antecedent’. We interpret  $\neg p$  again as obtained from  $\perp \rightarrow p$  by closed world reasoning. The completion then becomes  $(p \vee r \leftrightarrow q) \wedge (p \leftrightarrow \perp)$ , and reasoning from an interpretation is stuck: suppression. Indeed, in Byrne’s study [5] DA for this type of problem was applied by only 4% of the participants. However, in Dieussaert et al.’s study [11] 22% applied DA in this case. This could be a consequence of applying negation as failure to  $r$  as well, instead of only to abnormalities. The competence model allows both choices.

## 4.2 The backward inferences: MT and AC

As we have seen, the forward inferences rely on the completion, that is, the closed world assumption for facts. We propose that the backward inferences rely in addition on the closed world assumption for rules. When we come to discuss the neural implementation of the formalism we will see that this explains to some extent why backward inferences are perceived to be more difficult. This section uses the material on integrity constraints introduced in section 3.1.

**AC and MT for a single conditional premiss** Suppose we have a single conditional premiss  $p \wedge \neg ab \rightarrow q$  and a fact  $q$ . Closed world reasoning about facts would yield the completion  $\{((p \wedge \neg ab) \vee \top) \leftrightarrow q; ab \leftrightarrow \perp\}$ , from which nothing can be concluded about  $p$ .

But now assume that reasoning to an interpretation sets up the problem in such a way that AC is interpreted as an integrity constraint, that is, as the statement

if  $?q$  succeeds, then  $?p$  succeeds.

In this case, closed world reasoning for rules can be applied and we may ask what other atomic facts must hold if  $q$  holds. Since the only rule is  $p \wedge \neg ab \rightarrow q$ , it follows that  $p \wedge \neg ab$  must hold. For  $\neg ab$  this is guaranteed by closed world reasoning about facts, but the truth of  $p$  must be posited. In this sense AC is valid.

For MT, the reasoning pattern to be established is

if  $?q$  fails, then  $p$  fails.

One starts from the integrity constraint that the query  $?q$  must fail. This can only be if at least one of  $p$  and  $\neg ab$  fails. Since in this situation we know that  $\neg ab$  is true (by closed world reasoning for facts), we must posit that  $p$  is false.<sup>28</sup>

---

<sup>28</sup>This shows that subjects may do a *modus tollens* inference for the ‘wrong’, i.e. non-

**AC and MT for an additional premiss** In the case of an additional premiss, the program consists of

$$p \wedge \neg ab \rightarrow q, r \wedge \neg ab' \rightarrow q, \neg p \rightarrow ab', \neg r \rightarrow ab.$$

Consider AC: here we start with the integrity constraint that the query  $?q$  succeeds. It follows that at least one of  $p \wedge \neg ab, r \wedge \neg ab'$  must be true. But given the information about the abnormalities furnished by closed world reasoning for facts, namely  $\neg r \leftrightarrow ab$  and  $\neg p \leftrightarrow ab'$ , in both cases this means that  $p$  and  $r$  must true, so that AC is supported.

Now consider MT, for which we have to start from the integrity constraint that  $?q$  must fail. The same reasoning as above shows that at least one of  $p, r$  must fail – but we don't know which. We thus expect suppression of MT.

**AC and MT for an alternative premiss** In the case of AC, an alternative premiss leads to clear suppression (from 55% to 16%). It is easy to see why this must be so. Closed world reasoning for facts reduces the two conditional premisses to  $p \rightarrow q$  and  $r \rightarrow q$ . Given that  $?q$  must succeed, closed world reasoning for rules concludes that at least one of  $p, r$  must be true – but we don't know which. It is interesting at this point to look at an experiment in Dieussaert et al. [11] where subjects are allowed to give compound answers such as  $p \wedge q$  or  $p \vee q$ .<sup>29</sup> For AC, 90.7% subjects then chose the nonmonotonically correct  $p \vee r$ .

Turning to MT, we do not expect suppression, and indeed, the requirement that  $?q$  fails means that both  $p$  and  $r$  have to be false. In Dieussaert et al.'s data for MT, 96.3% of the subjects, when allowed to choose compound answers, chose the classically correct  $\neg p \wedge \neg r$ .

### 4.3 Competence and performance

We derived the pattern of suppression and non-suppression observed by Byrne and others by means of a nonclassical competence model, namely

---

classical, reason. There may be a relation here with the interesting observation that for children, the rate of MT seems to increase to around 74% for 10–12 year olds, only to decrease again for older children, to the usual percentage of around 50% (see [14] for discussion) Could it be that the younger children are good at MT for the wrong reason, and that the later decline is due to a not yet fully sufficient mastery of the classical semantics?

<sup>29</sup>These authors claim that Byrne's experiment is flawed in that a subject is allowed to judge *only for an atomic proposition or its negation* whether it follows from the premisses supplied. This would make it impossible for the subjects that draw a conclusion which pertains to both conditional premisses. 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 unfortunately not interpretable.

closed world reasoning with facts and rules, applied in the service of constructing a model of the discourse. We have thus adopted what Dieussaert et al. [11] call an ‘integration strategy’: the premisses, both conditionals and facts, are taken jointly when constructing a minimal model, and it is assumed that no more premisses will be supplied.

The data obtained by Dieussaert et al. [11] suggest however that some answers may not so much reflect a competence model as processing constraints. Dieussaert et al. observed that in the case of MT for an additional premiss, 35.3% of subjects, when allowed to choose compound answers, chose  $\neg p \wedge \neg r$ , whereas 56.9% chose the classically correct  $\neg p \vee \neg r$ . The second answer is readily explained in the present framework. The first answer can be explained if subjects first set  $ab$ ,  $ab'$  to  $\perp$ , and report the result of this intermediate computation: for if  $?q$  fails, both  $p \wedge \neg ab$  and  $r \wedge \neg ab'$  must be false. Since by assumption  $\neg ab$  and  $\neg ab'$  are true, both  $\neg p$  and  $\neg r$  must be true. We hypothesize that in a case such as this, some subjects simplify their handling of the abnormalities, to reduce the complexity of the correct derivation. This type of answer does not conform to a particular competence model, but reflects processing constraints. It is argued in [47] that these processing constraints play a role in autism as well.

#### 4.4 Summary

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 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; more importantly, in this language the conditional is a special, non-iterable non-truthfunctional connective. 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 example, in our case the definition of validity is given in terms of minimal models, leading to a nonmonotonic concept of validity. Once logical form is thus fixed (but not before!), one may inquire what follows from the premisses 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 contributes to the choice of a logical form

appropriate to content and context. 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; indeed default systems can provide several models of the variety of reasoning behaviour observed.

As regards Byrne’s positive suggestion that ‘mental models’ provides a better account of what goes in a subject who suppresses MP, we do not dispute that such subjects construct semantic representations (‘models’) which do not support MP. The present proposal seems however to add more detail as to how the construction is done, namely as a form of closed world reasoning analogous to what is necessary for planning.

We now propose that it is possible to go further and use these systems as a basis for proposals about cognitive architecture, in particular neural implementation.

## 5 Closed world reasoning and working memory

We will now show that the computations underlying the suppression effect can actually be performed very efficiently in suitable neural networks. The observation that there is a strong connection between logic programming and neural nets is not new (see d’Avila Garcez, Broda and Gabbay [2]), but what is new here is a very straightforward modelling of closed world reasoning (negation as failure) by means of coupled neural nets. This exploits the soundness and completeness of negation as failure with respect to Kleene’s strong three-valued logic. What is of importance here is that the relevant models can also be viewed as stable states of a neural network, obtained by a feedforward computation mimicking the action of the consequence operator associated to the logic program. We first present some pertinent definitions<sup>30</sup>.

### 5.1 Neural nets

**Definition 11** *A computational unit, or unit for short, is a function with the following input-output behaviour*

1. *inputs are delivered to the unit via links, which have weights  $w_j \in \mathbb{R}$*
2. *the inputs can be both excitatory or inhibitory; let  $x_1 \dots x_n \in \mathbb{R}$  be excitatory, and  $y_1 \dots y_m \in \mathbb{R}$  inhibitory*
3. *if one of the  $y_i$  fires, i.e.  $y_i \neq 0$ , the unit is shut off, and outputs 0*
4. *otherwise, the quantity  $\sum_{i=1}^{i=n} x_i w_i$  is computed; if this quantity is greater than or equal to a threshold  $\theta$ , the unit outputs 1, if not it outputs 0*

---

<sup>30</sup>For expository purposes we consider only very simple neurons, whose thresholds are numbers, instead of functions such as the sigmoid.

5. we assume that this computation takes one time-step.<sup>31</sup>

**Definition 12**

1. A spreading activation network is a directed graph on a set of units, whose (directed) edges are called links.
2. A (feedforward) neural network is a spreading activation network with two distinguished sets of units,  $I$  (input) and  $O$  (output), with the added condition that there is no path from a unit<sup>32</sup> in  $O$  to one in  $I$ .

Neural networks and spreading activation networks differ in some respects. A spreading activation network is typically conceived of as consisting of units which fire continuously, subject to a decay, whereas in neural nets one commonly considers units which fire once, when triggered. Some neural networks are equipped with a ‘nice’ structure, in order to facilitate the action of the backpropagation algorithm; e.g., one assumes that the network is composed of input, output and an ordered set of hidden layers, such that units are only connected to (all units of) the next layer. Spreading activation networks require no such assumption. Although we talk of ‘backpropagation’ in our spreading activation networks to describe a process analogous to the learning algorithm in neural nets, it does not have some of the ‘nasty’ neural implausibilities of the latter in this context. Our construction will need a bit of both kinds of network; in logic programming one typically computes models, which correspond to stable patterns of activation, but sometimes one focusses on models which make a designated output true, as when considering integrity constraints.

We now propose that the fixed points of the consequence operator associated to a program correspond to stable states in a spreading activation network derived from the program. We start again with positive programs; here the correspondence is due to d’Avila Garcez, Broda and Gabbay [2].

**5.1.1 Example**

Consider the language  $L = \{p, q, r, s, t\}$ , and following example of a program  $P$  in  $L$ :  $P = \{p, p \rightarrow q, q \wedge s \rightarrow r\}$ . We start from the empty model  $\mathcal{M}_0 = \emptyset$ , i.e. the model in which all proposition letters are false. We then get the following computation:

1.  $\mathcal{M}_1$  is given by  $T_P(\mathcal{M}_0)(p) = 1, T_P(\mathcal{M}_0)(q) = T_P(\mathcal{M}_0)(r) = T_P(\mathcal{M}_0)(s) = T_P(\mathcal{M}_0)(t) = 0$

---

<sup>31</sup>This assumption is customary for spreading activation networks and recurrent neural nets, although not for feedforward nets.

<sup>32</sup>The term ‘unit’ is unfortunate here, since strictly speaking input units do not compute anything; they are just nodes where data is fed into the network.

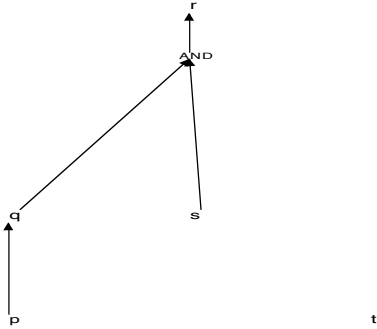


Figure 5: Network associated to  $\{p, p \rightarrow q, q \wedge s \rightarrow r\}$

2.  $\mathcal{M}_2$  is given by  $T_P(\mathcal{M}_1)(p) = 1$ ,  $T_P(\mathcal{M}_1)(q) = 1$ ,  $T_P(\mathcal{M}_1)(r) = T_P(\mathcal{M}_1)(s) = T_P(\mathcal{M}_1)(t) = 0$

The model  $\mathcal{M}_2$  is a fixed point of  $T_P$  in the sense that  $T_P(\mathcal{M}_2) = \mathcal{M}_2$ . It is also the least fixed point; we may consider  $\mathcal{M}_2$  to be the minimal model in the sense that it makes true as few proposition letters as possible.  $\mathcal{M}_2$  is not the only fixed point, since  $\mathcal{M}_3 = \{p, q, r, s\}$  is also one. However, such *greatest* fixed points are neurally less plausible; see below.

Figure 5 shows our proposal for the spreading activation network associated to the program  $P$ . Until further notice all links will be assumed to have weight 1. The important design decision here is that  $\rightarrow$  is not represented as a unit, but as a *link* – this is the neural correlate of the earlier observation that the conditional often does not act as a truthfunctional connective.

In this picture, the node labelled AND is a unit computing conjunction. The time-course of the spreading of activation in this network mimics the action of the monotone operator  $T_P$ . At time 0, no node is activated; at time 1, the data is fed into the network,<sup>33</sup> and only  $p$  is activated, and at time 2, as the result of a computation,  $p$  and  $q$  become activated. The presence of  $p$  in the program means that the input site  $p$  is continually activated, without decay of activation; the activation of  $q$  (and the nonactivation of  $r, s, t$ ), is therefore also permanent. The most important point to emphasise here is that the least fixed point of  $T_P$ , that is, the minimal model of  $P$ , corresponds to a stable pattern of activation of the network, starting from a state of no activation. We thus claim that the computation of minimal models is something that working memory is naturally equipped to do.

Interestingly, the greatest fixed point  $\mathcal{M}_3 = \{p, q, r, s\}$  cannot be obtained by a simple bottom up computation, starting from an initial state of no activation. In terms of neural nets, the greatest fixed point corresponds rather to starting from a state where all units are activated, and where the activation decays unless it is maintained by the input. This seems less

<sup>33</sup>Section 5.4 will contain a more detailed representation of input nodes.

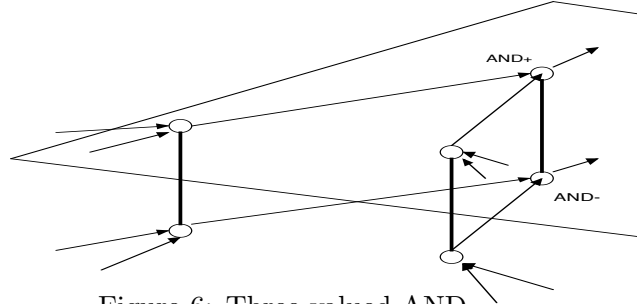


Figure 6: Three-valued AND

plausible as a model of neural computation.

## 5.2 From algorithm to neural implementation: definite programs

To describe a network corresponding to a definite logic program, we need units which compute  $\wedge$ ,  $\vee$  and  $\neg$  with respect to Kleene's strong three-valued logic. The following trick is instrumental in defining these units (here we will part company with [2]). The three truth values  $\{u, 0, 1\}$  in Kleene's logic can be represented as pairs  $(0, 0) = u$ ,  $(0, 1) = 0$  and  $(1, 0) = 1$ , ordered lexicographically via  $0 < 1$ . The pair  $(1, 1)$ , which would represent a contradiction, is considered to be excluded. We shall refer to the first component in the pair as the + (or 'true') component, and to the right component as the - (or 'false') component. Interpret a 1 neurally as 'activation', and 0 as 'no activation', so that  $u$  corresponds to no activation at all. Neural networks are now conceived of as consisting of two isomorphic coupled layers, one layer doing the computations for 'true', the other for 'false', and where the coupling is inhibitory to prevent pairs  $(1, 1)$  from occurring.

With this in mind, a three-valued binary AND can then be represented as a pair of units as in figure 6.

What we see here is two coupled neural nets, labelled + (above the separating sheet) and - (below the sheet). Each proposition letter is represented by a pair of units, one in the + net, and one in the - net. Each such pair will be called a *node*. The thick vertical lines indicate inhibitory connections between units in the + and - nets; the horizontal arrows represent excitatory connections. The threshold of the AND+ unit is 2, and that of the AND- unit is 1.

As an example, suppose the two truth values  $(1, 0)$  and  $(0, 0)$  are fed into the unit. The sum of the plus components is 1, hence AND+ does not fire. The sum of the - components is 0, so AND- likewise does not fire. The output is therefore  $(0, 0)$ , as it should be. There is an inhibitory link

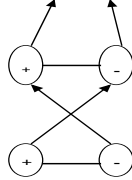


Figure 7: Three-valued negation

between the + and – units belonging to the same proposition letter (or logic gate) because we do not want the truth value (1, 1), i.e. both units firing simultaneously.

We obtain an AND-gate with  $n$  inputs for  $n \geq 2$  if the threshold for the plus-unit is set to  $n$ , and the threshold for the minus-unit is set to 1. Similarly, we obtain an OR-gate if the threshold of the plus-unit is set to 1, and that for the minus-unit to  $n$ . The reader may check that these conventions correspond to Kleene’s truth tables (see figure 4), reformulated in terms of pairs  $(i, j)$ . We also need a unit for negation, which is given in the figure 7, where each unit has threshold 1.

Now consider the logic program  $P$  we used to explain the suppression effect:

$$P = \{p, p \wedge \neg ab \rightarrow q, r \wedge \neg ab' \rightarrow q, \neg r \rightarrow ab, \neg p \rightarrow ab', \perp \rightarrow ab, \perp \rightarrow ab'\}.$$

For the sake of readability, in figure 8 we give only the + net; the diagram should be extended with a – net as in the diagram for AND (figure 6) In this picture, the links of the form  $0 \rightarrow ab$  represent the + part of the link from  $\perp$  to the pair of units corresponding to  $ab$ . A NOT written across a link indicates that the link passes through a node which reverses (1,0) and (0,1), and leaves (0,0) in place. AND indicates a three-valued conjunction as depicted above. The output node  $q$  implicitly contains an OR gate: its + threshold is 1, its – threshold equals the number of incoming links. The abnormality nodes likewise contain an implicit OR.

Before we examine how such networks could be set up in working memory, we trace the course of the computation of a stable state of this network, showing that  $q$  is not true in the minimal model of the program. Initially all nodes have activation (0,0). Then the input  $p$  is fed into the network, i.e. (1,0) is fed into the  $p$  node. This causes the  $ab'$  node to update its signal from (0,0) to (0,1), so that  $\neg ab'$  changes its signal to (1,0). But no further updates occur and a stable state has been reached, in which  $q$  outputs (0,0). If we view this state of activation as a (three-valued) model, we see that

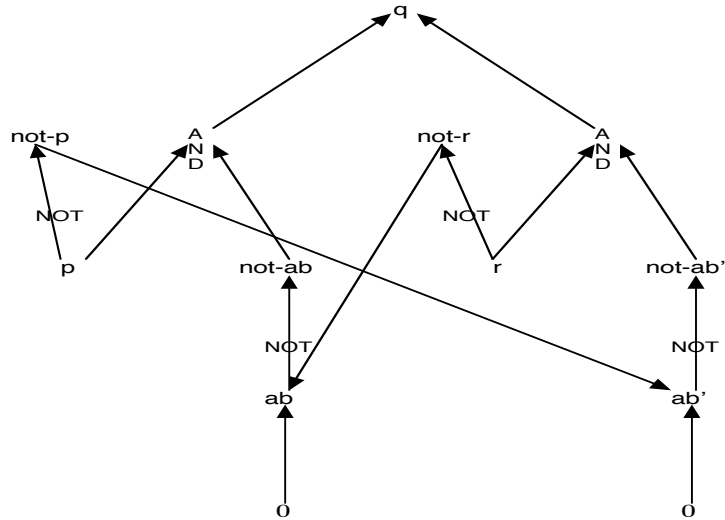


Figure 8: Network for the suppression of MP

$p$  is true, and the proposition letters  $r$  and  $q$  are undecided. Not surprisingly, this model is also the least fixed point of the three-valued consequence operator associated to the program.

### 5.3 Constructing the nets

We will now indicate briefly how such networks may be set up in working memory, following the highly suggestive treatment in a series of papers by Bienenstock and von der Malsburg [3, 49, 48]. Here we will only sketch a hypothesis to be developed more fully elsewhere. There are two issues to be distinguished: (1) the representation of conditionals as links, and (2) the structure of the network as a pair of isomorphic graphs. It seems that the work of Bienenstock and von der Malsburg is relevant to both issues.

Concerning the first issue, they observed that, apart from the ‘permanent’ connection strengths between nodes created during storage in declarative memory, one also needs variable connection strengths, which vary on the psychological time scale of large fractions of a second. The strength of these so-called dynamical links increases when the nodes which a link connects have the same state of activation; networks of this type are therefore described by a modified Hopfield equation. Applied to the suppression task, we get something like the following. Declarative memory, usually modelled

by some kind of spreading activation network, contains a node representing the concept ‘library’, with links to nodes representing concepts like ‘open’, ‘study’, ‘essay’ and ‘book’. These links have positive weights. Upon being presented with the conditional ‘if she has an essay, she will study late in the library’, these links become temporarily reinforced, and the system of nodes and links thereby becomes part of working memory, forming a network like the ones studied above. Working memory then computes the stable state of the network, and the state of the output node is passed on to the language production system. Modification of the connection strengths is an automatic (albeit in part probabilistic) process, and therefore the whole process, from reading the premisses to producing an answer, proceeds automatically. The second intriguing issue is how the two layers of neurons become isomorphically wired up through inhibitory interconnections. We believe this can be achieved using Bienenstock and van der Malsburg’s algorithm, since this is precisely concerned with establishing graph isomorphism. Unfortunately these brief indications must suffice here.

#### 5.4 Backward reasoning and closed world reasoning for rules

We now have to consider the computations that correspond to the inferences AC and MT. We have analyzed these by means of integrity constraints, that is, statements of the form ‘if query  $?φ$  succeeds/fails, then query  $?ψ$  succeeds/fails’. From a neural point of view, this is reminiscent of a form of backpropagation, except that in our case inputs, not weights are being updated. This distinction is however not absolute, and we will rephrase the construction in terms of the updating of weights. We propose that this difference between the processes of forward and backward reasoning can play a part in explaining various observations of the difficulty of backward reasoning such as MT. We will do only one example, AC for a single conditional premiss; this suffices to illustrate the main idea. We are given the premisses  $q$  and  $p \wedge \neg ab \rightarrow q$ . The corresponding network is given in figure 9.

In this network, if  $q$  becomes activated, then, provided nothing happens to maintain the activation, it will decay to a state of no activation. It follows that the only state of the network which will maintain the activation at  $q$  is one where  $p$  is (and remains) activated, and  $ab$  is not. One may rephrase this as an abductive learning problem: given an output (namely  $q$  active), the inputs have to be readjusted so that they yield the output. Now usually the input is given, and the weights in the network are readjusted. But since we are concerned with units which fire continuously, a reformulation in terms of the updating of weights is possible. For this purpose we replace what was previously an input node  $p$  by the configuration of figure 10.

Here, the bottom nodes always fire: the bold **1** indicates that this node stands for the true (or  $\top$ ); likewise **0** stands for the always false (or  $\perp$ ). Both nodes are taken to fire continuously. The links from the bottom nodes to

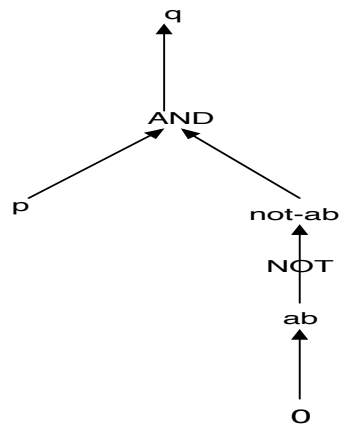


Figure 9: Network for AC

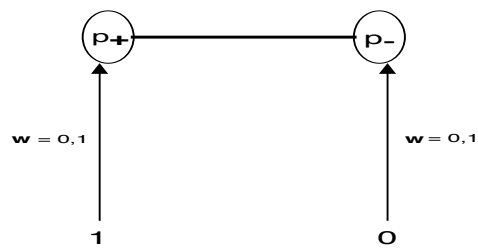


Figure 10: Structure of input nodes

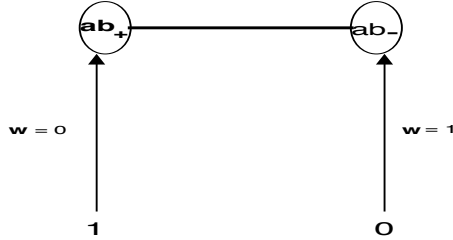


Figure 11: Structure of  $ab$  nodes

the  $p$ -nodes have weights 0 or 1. Activation of  $p$  with  $(1, 0)$  is now modelled as the left link having weight 1, and the right link having weight 0, and analogously for  $(0, 1)$ , with ‘left’ and ‘right’ interchanged. Nonactivation of  $p$  corresponds to both links having weight 0. In terms of this view of input nodes, the links previously described as  $\mathbf{0} \rightarrow ab$  now have the structure as in figure 11, where the weights are fixed.

Consider again the network of figure 9, and think of this as the + part of a net as in figure 6. Assume the inputs are as given by figures 10 and 11. In this network, the inputs always fire, and the output may, or may not, fire. Reaching a stable state now corresponds to readjusting the weights to match the observed input/output pattern. For instance, if  $q$  is activated, simple perceptron learning will update  $p$ ’s initial weight configuration (*left* : 0, *right* : 0) to (*left* : 1, *right* : 0).

In conclusion of this discussion of how the networks compute, we may thus note that the relatively easy forward reasoning patterns such as MP and DA correspond to simple feedforward networks, in which no learning occurs, whereas backward reasoning patterns such as AC and MT, which subjects typically find harder, necessitate the computational overhead of readjusting the network by an (albeit simple) form of backpropagation.

## 6 Discussion

The purpose of this paper has been to show that studies of reasoning, once taken beyond the ‘mental rules’ *versus* ‘mental models’ debate, may begin to address the important issue of the cognitive capabilities underpinning reasoning.

The first step consisted in clearing the ground, by showing that the imposition of logical form on natural language utterances involves a process of

parameter-setting far more complex than, say, associating a material implication to ‘if . . . then’. This process was dubbed ‘reasoning to an interpretation’; it was postulated to underlie credulous interpretation of discourse, in which the hearer tries to construct a model for the speaker’s discourse. We saw that the logical form assigned to the conditional introduces a parameter (namely *ab*), which plays an important role in *integrating* the conditionals with knowledge already present in declarative memory.

In the second step, we identified a logic that is a natural candidate for discourse integration: closed world reasoning, here treated formally as logic programming with negation as failure. Closed world reasoning is much easier on working memory than classical logic: a single model of the premisses suffices, and this model can be generated very efficiently. We showed that the phenomena observed in the suppression task conform to the pattern predicted by logic programming.

In the third step we showed that logic programming, unlike classical logic, has an appealing neural implementation. On our proposals working memory maintains a minimal preferred model of the world under description in a form in which at any point the model can be efficiently revised in the light of new information. This is the kind of system required for an organism to plan, and is the kind of system which might be developed by an organism evolved for planning which then turned its architecture to the performance of intentional communication.

Reflecting back on the psychology of reasoning literature, the computational properties of this logic are highly suggestive as candidates for ‘System 1 processes’ much discussed in dual process theories of reasoning (e.g. Pollock [28], Stanovich [37] and Evans [13])

System 1 is . . . a form of universal cognition shared between animals and humans. It is actually not a single system but a set of subsystems that operate with some autonomy. System 1 includes instinctive behaviours that are innately programmed and would include any innate input modules of the kind proposed by Fodor . . . The System 1 processes that are most often described, however, are those that are formed by associate learning of the kind produced by neural networks. . . . System 1 processes are rapid, parallel and automatic in nature; only their final product is posted in consciousness (Evans [13, p. 454]).

In these theories, logical reasoning is considered to belong to ‘System 2’ which is ‘slow and sequential in nature and makes use of the central working memory system [13, p. 454]’. Our proposals certainly challenge the idea that fast and automatic processes are thereby not logical processes, thus drawing a slightly different boundary between System 1 and System 2 processes.

Our proposals are for implementation in spreading activation networks and it is perhaps worth comparing the strengths and weaknesses of these to connectionist networks. In some ways spreading activation networks are

more immediately neurally plausible than connectionist networks. If activation of a node is interpreted as representing sustained firing of neurons, possibly with gradual decay of firing rate or intensity, then spreading activation networks more directly mimic neural activity than do connectionist networks. For the latter, it is not so clear how individual spikes or bursts of spikes are represented in the network simulations.

However, spreading activation networks are localist representations and are therefore not primitively capable of pattern completion, fault tolerance, etc. Connectionist networks are mainly designed for learning applications, but the process which constructs the networks proposed here, on the basis of discourse comprehension is not a conventional learning process. The process that most closely corresponds to learning in that construction is the ‘wiring’ between the pairs of nodes representing propositions, and the setting of thresholds to represent connectives. We have suggested how this process can be achieved in a neurally plausible fashion.

One last issue concerns a charge frequently brought against nonmonotonic reasoning, namely that its high computational complexity rules it out as a formal description of actual human reasoning. There are in fact two issues here, one pertaining to search for possible exceptions or abnormalities in the mental database, the other to the computational complexity of the derivability relation of a given nonmonotonic logic. The first issue can be illustrated by a quote from Politzer [27, p. 10]

Nonmonotonicity is highly difficult to manage by Artificial Intelligence systems because of the necessity of looking for possible exceptions through an entire database. What I have suggested is a some kind of reversal of the burden of proof for human cognition: at least for conditionals (but this could generalise) looking for exceptions is itself an exception because conditional information comes with an implicit guarantee of normality.

Translated to our formalism, the difficulty hinted at by Politzer concerns tabulating all clauses of the form  $\varphi \rightarrow ab$  which are present in memory. But here we do not have a knowledge base in the sense of AI, with its huge number of clauses which are all on an equal footing. The discussion of the neural implementation in section 5 has hopefully made clear that what counts is the number of links of the form  $\varphi \rightarrow ab$  which are activated in *working memory* by means of a mechanism such as Bienenstock and von der Malsburg’s ‘fast functional links’. This search space will be very much smaller. There remains the issue of how *relevant* information in long term memory is recruited into working memory, though we assume this is achieved efficiently through the organisation of long-term memory. We do not pretend to have solved the problem, but equally we do not believe the AI experience of its intractability is entirely relevant.

The second issue is concerned with the computational complexity of the decision problem for the relation ‘ $\psi$  is derivable from  $\varphi_1, \dots, \varphi_n$  in nonmono-

tonic logic  $\mathcal{L}'$ , where we may restrict attention to propositional logics only. For example, one well-known nonmonotonic logic, Reiter's *default logic* is computationally more complex than classical propositional logic, which is NP-complete, so in practice exponential (see Gottlob [16] for a sample of results in this area). By contrast, if  $\mathcal{L}$  is propositional logic programming with negation as failure, the corresponding decision problem is P-complete, hence *less* complex than classical propositional logic (see Dantsin et al. [10] and references given therein for discussion). This difference is mainly due to a restriction in the syntactic form of the rules, which have to be of the form  $\varphi \rightarrow A$ , where  $\varphi$  can be arbitrary, but  $A$  must be atomic. This restriction, whose main effect is to rule out disjunctions in the consequent, is harmless in the case of the suppression task. We do not deny that there may be other nonmonotonic reasoning tasks where this restriction causes problems; it should be noticed, however, that logic-programming has a certain track record in problem solving in AI, which provides further evidence of this logic's expressiveness.

## References

- [1] A. Athanasiadou and R. Dirven. *On conditionals again*. John Benjamins, Amsterdam, 1997.
- [2] A. d' Avila Garcez, K.B. Broda, and D. Gabbay. *Neural-symbolic learning systems: foundations and applications*. Springer, London, 2002.
- [3] E. Bienenstock and C. von der Malsburg. A neural network for invariant pattern recognition. *Europhysics Letters*, 4(1):121–126, 1987.
- [4] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation theoretic approach to default reasoning. *Artificial Intelligence*, 93(1–2):63–101, 1997.
- [5] R.M.J. Byrne. Suppressing valid inferences with conditionals. *Cognition*, 31:61–83, 1989.
- [6] R.M.J. Byrne, O. Espino, and C. Santamaria. Counterexamples and the suppression of inferences. *Journal of Memory and Language*, 40:347–373, 1999.
- [7] D. Chan and F. Chua. Suppression of valid inferences – syntactic views, mental models and relative salience. *Cognition*, 53(3):217–238, 1994.
- [8] A. Collins. Fragments of a theory of human plausible reasoning. In *Theoretical Issues in Natural Language Processing II*, pages 194–201. Urbana, IL: University of Illinois Press, 1978.
- [9] A. Collins and R. Michalski. The logic of plausible *Cognitive Science*, 13:1–49, 1989.
- [10] E. Dantsin, T. Eiter, G. Gottlob, and A. Voronkov. Complexity and expressive power of logic programming. *ACM Computing Surveys*, 33(3):374–425, 2001.

- [11] K. Dieussaert, W. Schaeken, W. Schroyen, and G. d'Ydewalle. Strategies during complex conditional inferences. *Thinking and reasoning*, 6(2):125–161, 2000.
- [12] K. Doets. *From logic to logic programming*. The M.I.T. Press, Cambridge, MA, 1994.
- [13] J.St.B.T. Evans. In two minds: dual-process accounts of reasoning. *TRENDS in Cognitive Sciences*, 7(10):454–459, 2003.
- [14] J.St.B.T. Evans, S.L. Newstead, and R.M. Byrne. *Human reasoning: the psychology of deduction*. Lawrence Erlbaum Associates, Hove, Sussex, 1993.
- [15] C. George. The endorsement of premisses – assumption-based or belief-based reasoning. *British Journal of Psychology*, 86:93–111, 1995.
- [16] G. Gottlob. Complexity results for nonmonotonic logics. *J. of Logic and Computation*, 2(3):397–425, 1992.
- [17] P.M. Greenfield. Language, tools and the brain: the ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and brain sciences*, 14:531–595, 1991.
- [18] P. Johnson-Laird. *Mental Models*. Cambridge University Press, Cambridge, 1983.
- [19] S. C. Kleene. *Introduction to Metamathematics*. North-Holland, Amsterdam, 1951.
- [20] R. A. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4:65–97, 1986.
- [21] R.A. Kowalski. Using meta-logic to reconcile reactive with rational agents. In *Meta-logics and logic programming*, pages 227–242. MIT Press, 1995.
- [22] A. Lechler. Interpretation of conditionals in the suppression task. Msc thesis, HCRC, University of Edinburgh., 2004.
- [23] J. McCarthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.
- [24] B. McGonigle, M. Chalmers, and A. Dickinson. Concurrent disjoint and reciprocal classification by *cebus apella* in serial ordering tasks: evidence for hierarchical organization. *Animal Cognition*, 6(3):185–197, 2003.
- [25] M. Oaksford and N. Chater. Probabilities and pragmatics in conditional inference: suppression and order effects. In D. Hardman and L. Macchi, editors, *Thinking: Psychological perspectives on reasoning, judgment and decision*, chapter 4, pages 95–122. Wiley, London, 2003.
- [26] F.J. Pelletier and R. Elio. What should default reasoning be, by default? *Computational Intelligence*, 13(2):165–187, 1997.
- [27] G. Politzer. Reasoning, judgment and pragmatics. In I.A. Noveck and D. Sperber, editors, *Experimental pragmatics*, chapter 4. Palgrave MacMillan, London, 2004.

- [28] J.L. Pollock. Oscar; a general theory of rationality. In J. Cummins and J.L. Pollock, editors, *Philosophy and AI: essays at the interface.*, pages 189–213. MIT Press, Cambridge MA, 1991.
- [29] J.L. Pollock. The logical foundations of goal-regression planning in autonomous agents. *Artificial Intelligence*, 106(4):267–335, 1998.
- [30] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47, 1988.
- [31] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
- [32] L.J. Rips. Cognitive processes in propositional reasoning. *Psychological Review*, 90:38–71, 1983.
- [33] M. Shanahan. Representing continuous change in the event calculus. In *Proceedings ECAI 90*, pages 598–603, 1990.
- [34] M. Shanahan. A circumscriptive calculus of events. *Artificial Intelligence*, 77:249–287, 1995.
- [35] M. Shanahan. Robotics and the common sense informatic situation. In *Working Notes of Common Sense 96, The Third Symposium on Logical Formalisation of Commonsense*, pages 186–198, 1996.
- [36] M.P. Shanahan. *Solving the frame problem*. The M.I.T. Press, Cambridge MA, 1997.
- [37] K.E. Stanovich. *Who is rational? Studies of individual differences in reasoning*. Lawrence Erlbaum, Mahwah, N.J., 1999.
- [38] M. Steedman. Plans, affordances and combinatory grammar. *Linguistics and Philosophy*, 25(5–6):725–753, 2002.
- [39] K. Stenning and R. Cox. Rethinking deductive tasks: relating interpretation and reasoning through individual differences. 2003. Submitted.
- [40] K. Stenning and J. Oberlander. A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cognitive Science*, 19:97–140, 1995.
- [41] K. Stenning and M. van Lambalgen. Semantics as a foundation for psychology. *Journal of Logic, Language, and Information*, 10(3):273–317, 2001.
- [42] K. Stenning and M. van Lambalgen. A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, 28(4):481–530, 2004.
- [43] K. Stenning and P. Yule. Image and language in human reasoning: a syllogistic illustration. *Cognitive Psychology*, 34:109–159, 1997.
- [44] R. Stevenson and D. Over. Deduction from uncertain premisses. *Quarterly Journal of Experimental Psychology A*, 48(3):613–643, 1995.
- [45] E. Traugott, A. ter Meulen, J.S. Reilly, and C.A. Ferguson. *On conditionals*. Cambridge University Press, Cambridge, 1982.
- [46] M. van Lambalgen and F. Hamm. *The proper treatment of events*. To appear with Blackwell Publishing, Oxford and Boston, 2004. Until publication, manuscript available at <http://staff.science.uva.nl/~michiell>.

- [47] M. van Lambalgen and H. Smid. Reasoning patterns in autism: rules and exceptions. In L.A. Perez Miranda and J.M. Larrazabal, editors, *Proc. Eighth International Colloquium on Cognitive Science Donostia/San Sebastian*. Kluwer, 2004.
- [48] C. von der Malsburg. Pattern recognition by labeled graph matching. *Neural networks*, 1:141–148, 1988.
- [49] C. von der Malsburg and E. Bienenstock. A neural network for the retrieval of superimposed connection patterns. *Europhysics Letters*, 3(11):1243–1249, 1987.

Role	Content
Conditional 1	If she has an essay to write she will study late in the library
Categorical	She has an essay to write
Conclusion	She will study late in the library (MP 88.3%)
Additional	If the library stays open, she will study late in the library.
Conclusion	She will study late in the library (MP 60.6%)
Alternative	If she has a textbook to read, she will study late in the library
Conclusion	She will study late in the library (MP 93.3%)
Conditional 1	If she has an essay to write she will study late in the library.
Categorical	She will study late in the library
Conclusion	She has an essay to write (AC 55.1%)
Additional	If the library stays open then she will study late in the library.
Conclusion	She has an essay to write (AC 53%)
Alternative	If she has a textbook to read, she will study late in the library.
Conclusion	She has an essay to write (AC 16%)
Conditional 1	If she has an essay to write she will study late in the library.
Categorical	She doesn't have an essay to write
Conclusion	She will not study late in the library (DA 49.3%)
Additional	If the library stays open, she will study late in the library
Conclusion	She will not study late in the library (DA 49.2%)
Alternative	If she has a textbook to read, she will study late in the library
Conclusion	She will not study late in the library (DA 22%)
Conditional 1	If she has an essay to write she will study late in the library.
Categorical	She will not study late in the library
Conclusion	She does not have an essay to write (MT 69.6%)
Additional	If the library stays open, she will study late in the library
Conclusion	She does not have an essay to write (MT 43.9%)
Alternative	If she has a textbook to read, she will study late in the library
Conclusion	She does not have an essay to write (MT 69.3%)

Table 1: Percentages of Dieussaert's subjects drawing target conclusions in each of the four argument forms *modus ponens* (MP), *modus tollens* (MT), *denial of the antecedent* (DA) and *affirmation of the consequent* (AC), in two premiss and three premiss arguments. 'Conditional 1' is the same first premiss in all cases. In a two-premiss argument it is combined only with the categorical premiss shown. In a three-premiss argument, both are combined with either an 'alternative' or an 'additional' conditional premiss.