Some Perspectives on Counterfactuals: Philosophical, Psychological and Computational

Diarmaid Ó Séaghdha
B.A. (Mod.) CSL
Final Year Project, May 3, 2003
Supervisor: Dr. Carl Vogel

May 3, 2003
Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university.

_____________________________ May 3, 2003
Diarmuid Ó Séaghda
Acknowledgements

Thanks to more people than I can mention. To my parents, without whom no counterfactual could have brought me to this point. To Carl, for 4 years of enthusiasm and support, and for many interesting ideas. To Julie, for keeping me sane. To Peter Keating, scholar and connectionist, for numerous late-night discussions about analyticity, aubergines, and the meaning of life. To great teachers, great friends and to 22 years of subtle intertwining influences.
If I were president of the USA I would make another planet.
First, I want another planet because it’s so boring to have 9 planets.
Next, I will get 5400 tons of rocks to build the new planet.
Last, I finished building the planet and the name is Vacation Island.

‘Dylan’, Mrs. Borelli’s Second Grade Class, Essrig Elementary School, USA
# Contents

1 **Introduction**  
1.1 Counterfactual Thoughts in Daily Life .......................... 2  
1.2 Linguistic Indicators of Counterfactuality ...................... 3  
1.3 Aims of the Dissertation ......................................... 5  

2 **Philosophical Perspectives**  
2.1 Some Problems with *If... Then*  ................................ 8  
2.2 ‘Metalinguistic’ Analyses ........................................... 15  
2.3 Possible Worlds Analyses .......................................... 17  
2.4 Partial Representations ............................................ 26  
2.5 Discussion .................................................................... 31  
  2.5.1 Factual and Counterfactual Conditionals .................... 31  
  2.5.2 Conversational Vagueness ...................................... 34  
  2.5.3 The Validity of Conditional Inference Rules ............... 39  
  2.5.4 Conclusion .......................................................... 44  

3 **Psychological Perspectives** ........................................ 45  
3.1 The Behavioural Psychology of Counterfactuals .................. 46  
  3.1.1 How to Mutate Reality ......................................... 46  
  3.1.2 Reasoning with Counterfactuals ............................... 51  
3.2 A Theory of Reasoning: Mental Logic .............................. 55  
3.3 Another Theory of Reasoning: Mental Models .................... 63  
3.4 Discussion .................................................................... 79  
  3.4.1 Models or Rules? .................................................. 79  
  3.4.2 Content, Counterfactuals and Conceptual Structure ....... 88  
  3.4.3 Conclusion .......................................................... 102
## CONTENTS

### 4 On Philosophy and Psychology

4.1 On Logic and Thought ................................. 106
4.2 On Bounded Rationality .............................. 107
4.3 On Modes of Representation ......................... 112
4.4 On Conversation ..................................... 117
4.5 On Truth ............................................ 118
4.6 Conclusion ........................................... 121

### 5 Epilogue: Computational Perspectives

5.1 Counterfactuals in Artificial Intelligence ............ 125
Abstract

Counterfactual assertions about non-actual states of affairs have long fascinated and perplexed both philosophers and psychologists. In recent years, theorists in both fields have attacked the problem of counterfactuals with significant success. In this study, relevant advances in philosophy and psychology are compared and contrasted. Despite their differing aims and methodologies, the problems encountered by both disciplines are found to be very similar. In particular, most approaches have failed to explain how the implicit background assumptions underlying a given counterfactual are identified. Counterfactuals are also considered from the perspective of artificial intelligence.
Chapter 1

Introduction
1.1 Counterfactual Thoughts in Daily Life

Counterfactual and hypothetical thought is pervasive. Humans regularly ponder the way the world might be, but is not. We like to tell fantastical stories about people that do not exist, and never will, who inhabit places that do not exist, and never will, and do things that nobody has ever done, nor ever will. We commonly consider how things might be if we were to win the lottery, buy a new car, or learn Chinese. When we plan future actions, we are well advised to envisage their possible outcomes. To illustrate this point (should it need illustration), I quote a recent news article about a report on the future of the British educational system, which contained the following comment by the chief executive of the National College for School Leadership:

“The scenarios outlined in the report are not a menu of distinct choices, but are more snapshots of possible futures which show school leaders that the contours of the future can be mapped, even though the path taken is not yet determined. ‘If we can imagine what might happen then, we can assess what needs to be done now.”

– BBC News Online, 28 March 2003

Considering the possible is thus of great importance in our daily lives. But not only do we commonly consider what is possible, we also entertain thoughts of the impossible. If, two days after I buy an expensive sports car, I crash and break my leg, I may think to myself that I could have avoided the accident, had I not chosen such a powerful car; alternatively, I may think that I should have driven more slowly, or kept my eyes on the road. My thoughts will depend on the facts of the scenario, even though I am thinking counterfactually – it may be that I had been considering a car that was equipped with better safety features, but ultimately chose the one that was more aesthetically pleasing; it may be that I often drive fast, or it may that some emergency had caused me to rush home on this occasion. If another party was involved in the crash, I may blame him, or I may blame myself, for not doing what he (or I) might have done to prevent it. Mentally constructing other, non-actualised scenarios may help me avoid a similar accident in the future.
Psychologists associate counterfactual thought with a wide range of cognitive phenomena, including creativity, the abilities to identify causal sequences and to learn from mistakes, judgements of accountability, fault and blame, and feelings of grief, regret, envy, indignation and frustration (Kahneman & Tversky 1982, among others, call these ‘counterfactual emotions’). Akatsuko (1997: 781) has even claimed that “counterfactual thinking is an essential necessity for human existence.” So it is not surprising that counterfactual thought has received attention in numerous subfields of the discipline, notably in social and cognitive psychology. As we shall soon see, philosophers and artificial intelligence researchers have also paid a great deal of attention to counterfactuals; the relationship between these heterogeneous perspectives will provide the main focus of the work at hand.

1.2 Linguistic Indicators of Counterfactual-ity

In English, as in most European languages, counterfactual thoughts are generally expressed by a combination of the conditional construction (*if*. . . *then*) and a marked verbform. The archaic subjunctive mood is commonly employed, but subjunctive forms are often identical to past indicatives, and there can be a large degree of uncertainty as to the proper form where they diverge (cf. *if I were rich* . . . versus *if I was rich* . . . ). Indeed, subjunctive conditionals need not always express counterfactuality\(^1\), and counterfactuals may appear without any of their usual identifying traits. The following examples showcase a range of canonical counterfactuals, as well as counterfactual-like non-counterfactuals, and counterfactuals masquerading as other kinds of sentences.

1. If I were in Paris, I would be in France.
2. If I had brought my umbrella, I wouldn’t be wet.

\(^1\)In an experiment, Thompson & Byrne (2002) found that only 52% of subjunctive conditionals appeared to receive a counterfactual interpretation. However, the context explicitly stated that the speaker was ignorant of the truth or falsity of antecedent and consequent, and we should therefore not be surprised that the counterfactual interpretation was suppressed in this instance. Reassuringly, the experimenters reached the verdict that “subjunctive conditionals were more likely than indicative conditionals to be interpreted counterfactually” (p. 1158).
(3) If the CIA had assassinated Hitler in 1937, World War II would never have happened.

(4) If cows lived underground, they would be fish.

(5) If you looked out the window, you would see an interesting game of tennis. (subjunctive, not counterfactual)

(6) If that’s the best you can do, then I’m Mahatma Gandhi. (indicative, counterfactual)

(7) No Hitler, no A-Bomb. (no conditional construction, counterfactual; due to Lewis 1973)

(1)–(4) are counterfactual in both form and content, though they clearly differ in truth and acceptability. One would not hesitate in judging (1) to be true, given that world geopolitics are as they are; (2) seems unproblematic, assuming an appropriate context (an umbrella would not stop you getting wet if you were swimming); (3) is highly debatable, and its truth or falsity is unknowable – different individuals will presumably make different truth judgements based on their prior beliefs; (4) appears wholly false, yet this counterfactual was overheard in a genuine conversation – it is clear that a highly atypical discourse context is required to render it felicitous. Example (5) is a subjunctive conditional like the others, but it does not express counterfactuality; rather it expresses the same thing as the corresponding indicative conditional If you look outside, you will see an interesting game of tennis, i.e. a factual possibility. (6) is syntactically indicative, but its most probable reading is a counterfactual one (assuming the speaker is not Mahatma Gandhi), indicating that the speaker believes the hearer is not doing the best he/she can. The issues posed by this sentence can be readily explained in terms of Gricean maxims, but the example suffices to show that subjunctive conditionals are not the only way of signalling counterfactuality. (7) also shows this; there is no if...then, merely an association of two entities, but a hearer with appropriate world knowledge should elicit a counterfactual interpretation.

2For example, an individual may know that the CIA was not founded until 1947, and may attribute a high degree of salience to this fact.

3It may well be that (5) conveys a higher degree of hypotheticality (which may correlate with a pragmatic function of signalling politeness or formality) than its indicative counterpart. Similar considerations lead Comrie (1986) to reject the idea that the ‘counterfactual construction’ can be identified with a counterfactual meaning, merely with a certain degree of hypotheticality.
CHAPTER 1. INTRODUCTION

Some languages, like Mandarin and Indonesian, possess no means of expressing different degrees of probability, and thus do not distinguish syntactically between assertions about the future, factual conditionals, and counterfactuals. This has led some (notably Bloom 1981) to the Whorfian conclusion that speakers of these languages are less adept at hypothetical thought than speakers of other, more flexible, languages. Such a controversial thesis was bound to be challenged, and subsequent studies (e.g. Liu 1985) have contradicted Bloom’s findings, as well as identifying serious methodological problems in his experiments. It seems safe to continue to maintain that counterfactual thought is an important cognitive process in all humans.

For the purposes of this study, the term ‘counterfactual’ will be identified with the archetypal subjunctive conditional whose speaker knows or believes both antecedent and consequent to be false. This move is justifiable on a number of grounds: firstly, the meanings of apparent counterexamples like (5),(6) and (7) can be explained in pragmatic terms; secondly, our main concern is with counterfactual meanings, not surface indicators of counterfactuality.

1.3 Aims of the Dissertation

My primary aim in this work is to bring together the insights of philosophers and psychologists who have studied counterfactuals, to identify the common concerns of the two disciplines, and to identify where those concerns diverge. For a philosophical account of the semantics of counterfactuals should not contradict the behavioural facts of how people use them, and any psychological theory should be of some philosophical significance, if it is to be of any general interest. I shall also briefly consider how these considerations can be integrated with issues in computer science. The investigation is not really aligned to any recognised field of scholarship, although it may claim to be an essay in cognitive science.

In Chapter 2 I describe some of the most important philosophical theories of counterfactuals. I then discuss three of the main issues raised by these theories: the desirability of a unified theory of conditional and counterfactual semantics; the problem of vagueness in language; and the failure of certain ‘canonical’ inference rules. In Chapter 3 I give an account of psychological experiments that have studied the generation of, and reasoning with, counterfactual thoughts. I then describe two of the foremost psychological theories
of reasoning, and discuss their strengths and shortcomings. Special attention is devoted to the effects of content and world knowledge on counterfactual reasoning. In Chapter 4 I attempt to unify the main themes of the preceding chapters and to reach some conclusions about how people judge the truth or falsity of a given counterfactual. In Chapter 5 I (very) briefly outline the study of counterfactuals in artificial intelligence research.
Chapter 2

Philosophical Perspectives
CHAPTER 2. PHILOSOPHICAL PERSPECTIVES

2.1 Some Problems with *If*... *Then*

“Even the ravens on the roof tops are croaking about which conditionals are true.”
– Callimachus, quoted in Lukasiewicz (1970)

Philosophical discussions of counterfactuals coincide with the ancient debate regarding the semantics of conditionals in general. The Stoic philosopher Diodorus Cronus (died ca. 307 BC) argued with his pupil Philo of Megara about the meaning of the statement *If it is day, then the sun is shining* (Lukasiewicz 1970). Philo argued that such a conditional is true if both its antecedent and its consequent are true, or if its antecedent is false. Diodorus, on the other hand, held that a conditional is true if and only if it is not possible, and never has been, that the antecedent be true and the consequent false.

Modern logic since Frege has adopted the Philonian, or material, conditional as the standard logical counterpart of *if*... *then*. Like the other logical connectives, the material conditional is truth-functional – the truth value of *if* *p*, then *q* is wholly determined by the truth values of *p* and *q* and the ‘meaning’ of *if*, as given by the familiar truth table for → (Table 2.1).

<table>
<thead>
<tr>
<th><em>p</em></th>
<th><em>q</em></th>
<th><em>p → q</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Table 2.1: The standard truth table for →

However, the truth-functional analysis has not been an unqualified success. The *if*... *then* construction is far more flexible than the more well-behaved connectives like *and* and *or*, and is thus less amenable to a reductionist semantics. As many philosophers and linguists have observed, the truth table for → often fails to correspond with the way conditionals are used in natural language. For example, the material conditional licenses the

1Though C.I. Lewis favoured a ‘strict implication’ interpretation closer to that of another Stoic philosopher, Chrysippus (cf. Lewis & Langford 1932).
inferences $\neg p \vdash p \to q$, and $q \vdash p \to q$, i.e. any conditional with a false antecedent or a true consequent is automatically true. Thus, the following conditionals are all (in the case of (9), at least contingently) true:

(8) a. If 20 is a prime number, then water is blue.
   b. If 20 is a prime number, then water is not blue.

(9) a. If the Pope is a Scientologist, then Paris is the capital of France.
   b. If the Pope is a Catholic, then Paris is the capital of France.

Having to accept these conditionals as true seems wildly counterintuitive; accordingly, the inferences that license them are sometimes called the ‘paradoxes of material implication.’ A further ‘paradox’ that results from these is that $(p \to q) \lor (q \to p)$, i.e. that for any two propositions, one of them implies the other. It has been suggested (by, e.g., Quine 1972) that the meaning of if...then corresponds to a ‘defective’ truth table, so that if $p$, then $q$ has no truth value when $p$ is false; however, this in turn has the undesirable consequence that the biconditional $p$ if and only if $q$ can no longer be interpreted as if $p$ then $q$, and if $q$ then $p$ (Johnson-Laird & Byrne 2002).

A related problem is that the material conditional interpretation is blind to the content of the antecedent and consequent. Sentences like (9) seem doubly odd, because there is no clear connection between the religion of the Pope and the geopolitics of France.\(^2\) Even if we were to adopt Quine’s defective truth table, we would still be forced to concede that any conditional with true antecedent and consequent must itself be true.

Stevenson (1970: 28) observed that the following argument is valid under the material conditional interpretation: “This is false: if God exists then the prayers of evil men will be answered. So we may conclude that God exists, and (as a bonus) we may conclude that the prayers of evil men will not be answered.” For the negation of a material conditional is logically equivalent to the conjunction of its antecedent and the negation of its consequent. The apparent absurdity of the argument may follow from the fact that individuals usually deny the truth of a conditional by asserting the corresponding conditional with negated consequent, i.e. if $p$ then not $q$, and not by negating the entire conditional.

Yet it would be false to conclude that the material conditional interpretation of if...then is irrevocably flawed. In fact, many philosophers have

---

\(^2\)A study by Fillenbaum (1975) found that over 90% of subjects considered conditionals with unrelated antecedent and consequent to be “extraordinary or strange”.
defended this interpretation. If *if* is to be given a truth-functional semantics, then the material conditional is the only viable contender. Furthermore, there are many cases in which natural language conditionals do behave like truth-functional $\rightarrow$. The following inferences are valid in both logic and everyday linguistic usage:

1. **Modus Ponens**
   
   $p \rightarrow q$  
   If Jim has studied, he will pass the exam  

   $p$  
   Jim has studied  

   $\therefore q$  
   $\therefore$ He will pass the exam

2. **Contraposition**
   
   $p \rightarrow q$  
   If aubergines are obviously food, then the monkeys will eat them  

   $\therefore \neg q \rightarrow \neg p$  
   $\therefore$ If the monkeys will not eat them, then they are not obviously food

3. **Strengthening the Antecedent**
   
   $p \rightarrow q$  
   If Fianna Fáil win the next election, I will be very surprised  

   $p \land r \rightarrow q$  
   $\therefore$ If Fianna Fáil win the next election and it rains on election day, I will be very surprised

4. **Hypothetical Syllogism (Transitivity)**
   
   $p \rightarrow q$  
   If Jarrod knows some German, he will have no problems ordering a Pils  

   $q \rightarrow r$  
   If he has no problems ordering a Pils, he will order many glasses of Pils  

   $\therefore p \rightarrow r$  
   $\therefore$ If Jarrod knows some German, he will order many glasses of Pils

Strawson (1986) identifies *if* $p$, then $q$ as equivalent to $p \rightarrow q$, and claims that counterexamples such as (8) and (9) can be rebutted by invoking Grice’s (1975) conversational maxims. If one knows that $p$ is false, or that $q$ is true, then asserting *if* $p$, then $q$ violates the Maxim of Quantity, as it is less informative than asserting $\neg p$ or $q$ alone. It follows that one should only
assert if \( p \), then \( q \) when one has reason to believe that \( \neg(p \land \neg q) \) but is unaware of the truth or falsity of \( p \) and \( q \). This state of affairs is only likely to come about when one is aware of some connection (‘ground-consequent relation’) between \( p \) and \( q \), possibly based on logical principles, possibly based on knowledge about causal laws, social contracts, world geopolitics, etc. In accordance with Grice’s Cooperative Principle, the hearer may thus assume that the speaker of a conditional utterance does not know whether \( p \) or \( q \) are true or false, but has good reason to believe that some connection obtains such that \( p \land \neg q \) is false\(^3\). In this manner, we can distinguish between semantic truth conditions and extralogical assertability conditions that are governed by pragmatic rules and the state of discourse.

Lewis (1986d) also defends a material conditional interpretation of the indicative conditional. After a suggestion that he attributes to Frank Jackson, he proposes that the if...then construction conveys a conventional implicature of robustness, i.e. that the speaker’s subjective probability of the consequent and the speaker’s subjective probability of the consequent conditional on the antecedent\(^4\) are close together, and both are high, so that if one were at some point to learn that the antecedent were true, one would continue to find the antecedent probable. It is suggested that we sometimes value robustness over informativeness, for example when we believe that the hearer may have information that would falsify our stronger beliefs. Lewis observes that an analysis like Strawson’s cannot explain the acceptability of *Fred will not study, and if he does he still won’t pass* (p. 154). Gricean maxims dictate that if one knows \( \neg A \), one should not assert \( A \rightarrow B \); if we take the conditional assertion to signal robustness, then we explain that even if the prediction that Fred will not study turns out to be wrong, one still wishes to maintain that he won’t pass. A further consequence of Lewis’ theory is that conditionals which are true solely in virtue of the truth of their consequents are also assertable\(^5\). Support for this prediction comes from the existence of

\(^3\)A similar analysis is given in Gamut (1991: ch. 6).
\(^4\)The probability of \( A \) conditional on \( B \), \( P(A|B) = \frac{P(A\land B)}{P(B)} \).
\(^5\)The truth of all conditionals with true antecedents and true consequents is accepted in most logical frameworks. One exception is relevance logic, which as its name implies requires that the antecedent be relevant to the consequent for a conditional to be true (Read 1995). Relevance logic denies the automatic truth of both true antecedent-true consequent conditionals and conditionals with contradictory antecedents; the following conditionals are both predicted to be false:

(1) If second order logic is undecidable, then it is incomplete.
conditionals whose truth does just depend on the truth of the consequent, such as (10).

(10) If you’re hungry, I’m cooking a stew.

The inference rules given in 1–4 were introduced as evidence for equating if with the material conditional; however, counterexamples have been presented for each rule, claiming to show that these inferences are not always valid for natural language conditionals. For example, the following scenario from Adams (1965: 166) illustrates a failure of transitivity:

(11) If Brown wins the election, Smith will retire to private life. If Smith dies before the election, Brown will win it. Therefore, if Smith dies before the election, then he will retire to private life.

Violations of contraposition are not hard to come by; conditionals such as (10), whose truth/falsity derives from the truth/falsity of their consequent, do not generally license contraposition:

(10') If I’m not cooking a stew, then you’re not hungry.

Even the axiomatic Modus Ponens has been called into question; Lycan (1994: 223) reports an example of Allan Gibbard’s, which seems to contradict the universal validity of this rule:

(12) I’ll be polite if you insult me, but I won’t be polite if you insult my wife.

\[ \text{<The Hearer insults both the speaker and his wife>} \]

CONTRADICTION!!

Whether these are in fact true counterexamples to our proposed rules of inference is a contentious issue, which will arise again in our discussion of counterfactuals (see section 2.5). It may be that the inference rules are valid, and that apparent counterexamples arise when some restriction on implicit

---

(2) If Socrates is a man and not a man, then he is a stone.

(1) is false because undecidability does not entail incompleteness (first order logic is undecidable but complete); merely looking at the actual world or most similar antecedent-worlds is not enough to establish truth. That (2) be judged false requires allowing that some contradictions can be true – relevance logic is therefore a paraconsistent logic (cf. Priest 1998).
‘background conditions’ is violated; on the other hand, it may be that the inference rules are not valid, and merely tend to correspond with the world and how we describe it.

Ernest Adams (1965, 1975) adopts the latter explanation, and presents an analysis whereby truth-conditionality is superseded by considerations of probabilistic soundness. For an inference rule to be probabilistically sound, “it should be impossible for the premises of an inference to be probable while its conclusion is improbable” (Adams 1975: 1). Adams thus deals with the probability of propositions, which he equates with the probability that the proposition is true. The probability of a conditional is then equated with the corresponding conditional probability. It follows that Contraposition, Antecedent Strengthening and Transitivity are all probabilistically unsound, but tend to be useful in most situations. For example, Contraposition fails when the probability of the antecedent is too small, i.e. when a conditional is probable in virtue of the probability of its consequent.

There are a number of advantages to the probabilistic account. Firstly, we usually must reason from uncertain premises, and we generally hold some premises to be more dependable than others; bivalent logic cannot describe such comparative probability judgements. Secondly, it explains why logically sound inference rules sometimes fail. Thirdly, probabilistic analyses have been successful in accounting for other problematic linguistic phenomena, such as generic sentences (Cohen 1999). Yet despite these considerations, Adams’ theory has not been assimilated into the philosophical mainstream (though Edgington 1995 has recently defended a similar theory). One problem for Adams arises from Lewis’s (1986d) Triviality Result, which proves that the probability of a conditional (i.e. the probability that it is true) cannot be equated with the corresponding conditional probability for any non-trivial distribution of probabilities (i.e. for any probability function that assigns positive probability to more than two incompatible alternatives). Adams (1975) responds to this by introducing a partial probability function, so that the probability of a proposition is not always defined. Further problems arise when we consider the relationship of conditional sentences to the broader semantic framework. If the ‘meaning’ of a conditional sentence corresponds to a probability, then it is unclear how conditionals should interact.

---

6This comment is perhaps unfair, as it may give the impression that some rival theory has been adopted as the ‘mainstream’ theory.

7First published in 1976.
with other connectives, unless they too are to be interpreted probabilistically. Adams (1975) admits that his account cannot describe nested or conjoined conditionals. It seems unparsimonious to reject the fruits of traditional semantic theory, merely to accommodate a specific theory of conditionals, when the superiority of that theory seems unclear.

Whichever account we adopt for indicative conditionals, the corresponding problems multiply when we turn to face counterfactual conditionals. Interpreting counterfactual \( \text{if} \ldots \text{then} \) as the material conditional is clearly a non-starter; as both the antecedent and the consequent of a counterfactual are by definition false, it would follow from the truth table in Table 2.1 that all counterfactuals are vacuously true. Counterfactuals are not automatically true when their consequents are true, cf. *If Great Britain were a republic, Great Britain would have a monarchy*, unless a consequent is necessarily true. And whereas we can often judge the truth or falsity of an indicative conditional by observing the facts of the matter, there will be no observable facts in the actual world that by themselves enable us to adjudicate on counterfactuals. How, then, can we say that a given counterfactual is true or false (as we would clearly like to do)?

Ramsey (1931) made the following famous suggestion:

> If two people are arguing ‘If \( p \) will \( q \)?’ and are both in doubt as to \( p \), they are adding \( p \) hypothetically to their stock of knowledge and arguing on that basis about \( q \); so that in a sense ‘If \( p, q \)’ and ‘If \( p, \bar{q} \)’ are contradictories. We can say they are fixing their degrees of belief in \( q \) given \( p \). If \( p \) turns out false, these degrees of belief are rendered *void*. If either party believes \( \bar{p} \) for certain, the question ceases to mean anything to him except as a question about what follows from certain laws or hypotheses.’

(p. 247, footnote 1)

This footnote has exerted an influence out of all proportion to its length or status in the original text, to the extent that philosophers argue over whose theory most faithfully interprets the so-called ‘Ramsey Test’ (e.g. the debate in Read 1995 and Edgington 1995). This original formulation of the Ramsey Test is couched in explicitly epistemic terms (‘stock of knowledge’, ‘degrees of belief’), and it mentions the fact (observed earlier) that individuals deny the truth of a conditional by asserting the conditional *If \( p \), then not \( q \)*; in
these ways, it departs from the traditional logical analysis. In the case of counterfactual conditionals, we are discussing “what follows from certain laws or hypotheses.” If this brief explanation appears vague, we shall soon see that the subsequent 70 years of study have failed to fully resolve the vagueness.

2.2 ‘Metalinguistic’ Analyses

It was in the 1940s that philosophers first began to recognise the difficulties posed by counterfactuals in their own right. The approach adopted by these earliest theorists (Chisholm 1946, Goodman 1947) has come to be known as the ‘metalinguistic’ analysis. Counterfactuals were recognised to be of interest as topics in the philosophy of language, but more importantly, it was held that the ‘problem of counterfactuals’ “constitutes the basic problem in the logic of science” (Chisholm 1946: 302), due to its intimate relationship with the problem of distinguishing true contingent generalisations from genuine laws.

The core thesis of the metalinguistic account can be summarised as follows: counterfactual if φ then ψ is true iff ψ is derivable from φ ∧ Λ, where Λ is some suitable set of true statements. So the truth of (13) depends on the existence of an expanded conditional (14) whose antecedent entails the consequent.

(13) If that match had been scratched, it would have lighted. (Goodman 1947)
(14) If that match had been scratched, and the match was well made, and it was dry enough, and there was enough oxygen,...and every well-made, sufficiently dry match that is struck in sufficient oxygen...lights, it would have lighted.

We can identify two distinct kinds of ‘true statements’ that can determine the truth or falsity of a conditional. Firstly, relevant facts about the world – whether the match is well made, whether there is enough oxygen – must be considered; in many cases we cannot know these facts for certain and must assume that a situation is ‘normal’ unless we discover evidence to the contrary.

---

8This term is due to David Lewis (1973). As Hansson (1995) observes, there is nothing especially ‘linguistic’ about this approach; he prefers to describe it as the ‘derivability’ analysis. In any case, debates over appropriate nomenclature hardly affect the substance of the theory.
CHAPTER 2. PHILOSOPHICAL PERSPECTIVES

contrary, in other cases we can verify facts (to a high degree of certainty) by observing the world. Secondly, we must invoke laws of nature such as state that a well-made, sufficiently dry match, when stuck in an atmosphere of sufficient oxygen, etc. will light. These laws provide the ‘necessary connection’ that license the entailments which allow us to judge the truth or otherwise of a given counterfactual.

How can we know which ‘true statements’ are to be invoked? We clearly cannot answer: those true statements, should they exist, such as render the counterfactual in question true; for this is surely begging the question. Nor is it feasible to consider the set of all true statements, as this set will contain the negation of the antecedent. Nor may we consider any set of facts which, taken together, entail the negation of the antecedent. Goodman (1947) introduced the criterion of cotenability: \( A \) is cotenable with \( S \), and \( A \land S \) self-cotenable, if it is not the case that \( S \) would not be true if \( A \) were. We can thus reformulate the metalinguistic thesis as follows: counterfactual \( \text{if } \phi \text{ then } \psi \) is true iff \( \psi \) is derivable from \( \phi \land \Lambda \), where \( \Lambda \) is some suitable set of cotenable true statements. But as Goodman himself noted, the definition of cotenability itself contains a counterfactual, and he ultimately despaired that “we can never explain a counterfactual except in terms of others, so that the problem of counterfactuals must remain unsolved” (p.16). From a computational point of view, we can see that testing for cotenability is equivalent to testing for consistency in a knowledge base, a task that is recognised to be intractable. As we shall see in the next section, Lewis (1973) offered a definition of cotenability in terms of possible worlds.

Even should we arrive at a satisfactory account of cotenability, we still have no way of predicting which ‘true statements’ will be required to evaluate a given counterfactual. One problem facing us is exemplified by Chisholm’s Apollo sentences:

\[
\begin{align*}
(15) & \quad \text{a. If Apollo were a man, he would be mortal.} \\
& \quad \text{b. If Apollo were a man, at least one man would be immortal.}
\end{align*}
\]

(Chisholm 1946: 303)

The antecedent \textit{Apollo is a man} conflicts with a large number of interconnected prior beliefs about gods and men and moon rockets and many other things; it is a non-trivial task to selectively retain some and abandon others, as we must do to preserve cotenability. What the sentences in (15) demonstrate is that different counterfactuals with the same antecedent may require
different sets of statements – (15a) requires that we abandon our belief in Apollo’s immortality, and retain our belief in the generalisation that all men are mortal; (15a) requires that we do the direct opposite. At this point one tends to invoke notions of ‘context’ or ‘pragmatic vagueness’; that pragmatic factors are involved is beyond doubt, and such factors will be discussed further in section 2.5.

2.3 Possible Worlds Analyses

The concept of possible worlds is an old one, dating back at least as far as Leibniz, who famously asserted that we live in the best one of all. The usefulness of possible worlds in logical analysis, particularly in providing a semantics for modal logic, was explored by many philosophers in the twentieth century. Stalnaker (1968) provided an analysis of conditionals that builds upon the possible-worlds semantics developed by Kripke (1963).

Stalnaker’s starting point is Ramsey’s (1931) test for conditionals. He denies the need to posit a ‘necessary connection’ between antecedent and consequent, in view of conditionals whose truth derives from the truth of the consequent alone (Stalnaker here considers semifactuals, or even if-conditionals). Where necessary connections do play a role is in causing the (causal, logical, ...) consequences of a belief to be added to one’s ‘stock of belief’ when that belief is itself added; “since the rational man accepts the consequences of his beliefs” (p. 101). If the antecedent does not enter into any relation with the consequent, the truth of the conditional will depend on whether the consequent already belongs to the evaluator’s stock of beliefs. In the case of counterfactuals, Stalnaker interprets the Ramsey Test as follows:

First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true. (p. 102)

This seems to go beyond Ramsey’s original formulation, but no matter. Stalnaker justifies the move to a possible worlds analysis by claiming that “a possible world is the ontological analogue of a stock of hypothetical beliefs” (p. 102). This assertion is more controversial, but we shall accept it for the moment. The essential idea of Stalnaker’s theory (and of all possible worlds
analyses) is that we identify some alternative state of affairs (or some set of states of affairs) that is similar to our own, but in which the antecedent of the counterfactual under consideration holds, and we decide whether the consequent holds in that state of affairs. The concept of a possible world is equivalent to that of an ‘alternative state of affairs.’

A Stalnaker model for counterfactuals is a quintuple $< I, R, f, \lambda, [\ ] >$, where:

- $I$ is the set of all possible worlds.
- $R$ is a binary accessibility relation on $I$.
- $f$ is a selection function which takes a proposition and an element of $I$ to an element of $I$ where the proposition is true.
- $\lambda$ is a special element of $I$, the absurd world, in which contradictions and all their consequences are true. It is not accessible from any other member of $I$.
- $[\ ]$ is an interpretation which maps sentences onto subsets of $I$; for any sentence $\phi$, $[\phi] = \{ i | \phi \text{ is true at } i \}$.

A conditional connective $>$ can then be defined as follows:

- $A > B$ is true at world $\alpha$ iff $B$ is true at $f(A, \alpha)$.
- $A > B$ is false at world $\alpha$ iff $B$ is false at $f(A, \alpha)$.

So the bulk of the work is left to the selection function $f$. This function incorporates two separate functionalities: it is a selection function in the traditional sense, in that it selects a world where a given proposition is true; furthermore, it imposes a similarity constraint – the world that it selects should be that world which most resembles the input world. Some restrictions seem natural on the operations of $f$; for all antecedents $A$, $A'$, and base worlds $\alpha$:

1. $A$ must be true in $f(A, \alpha)$.
2. $f(A, \alpha) = \lambda$ only if there is no world possible w.r.t. $\alpha$ in which $A$ is true.
3. If \( A \) is true in \( \alpha \), then \( f(A, \alpha) = \alpha \).

4. If \( A \) is true in \( f(A', \alpha) \) and \( A' \) is true in \( f(A, \alpha) \), then \( f(A, \alpha) = f(A', \alpha) \).

(3) states that the world most similar to any world is itself; if the antecedent is true in the ‘base world’ (the world where the counterfactual is to be evaluated, usually the ‘actual’ world, that in which it is uttered), then we should test for the consequent in that world. (4) requires that the similarity metric adopted be ordered. In fact, it is implicit in the specification of \( f \) that for each base world there is a total ordering of worlds with regard to similarity, and that for any given antecedent, there will be a unique antecedent-world that is more similar to the base world than any other worlds (Stalnaker’s Uniqueness Assumption).

The purpose of the absurd world is to make vacuously true any counterfactual with an impossible antecedent, e.g. If 3 were 5, then roses would be violet. As usual, we can explain the unattractiveness of such truths by deeming them ‘inassertable’.

Stalnaker presents a formal system \( \textbf{C2} \) for his conditional logic; however, constraints on space prevent me describing it here. Modus Ponens and Modus Tollens are valid in \( \textbf{C2} \); Contraposition, Antecedent Strengthening and the Hypothetical Syllogism all fail\(^9\). One of the more controversial aspects of Stalnaker’s logic is its endorsement of the Law of Conditional Excluded Middle: \( (A > B) \lor (A > \neg B) \). We can read this as If \( A \), then either \( B \) or not \( B \); under the Uniqueness Assumption, the selection function will supply us with a single \( A \)-world, and given conventional Excluded Middle, \( B \) will be either true or false in that world. Some commentators, notably Lewis (1973) have objected to Conditional Excluded Middle, on the basis that it prevents one asserting sentences like (16).

(16) It is not the case that if Bizet and Verdi were compatriots, Bizet would be Italian; and it is not the case that if Bizet and Verdi were compatriots, Bizet would not be Italian; nevertheless, if Bizet and Verdi were compatriots, Bizet either would or would not be Italian. (Lewis 1973: 80)

\(^9\)That Modus Tollens is valid but Contraposition is not represents a notable feature of both Stalnaker and Lewis’ theories. It seems to follow from the further rejection of the Hypothetical Syllogism.
Whether one might wish to assert such a thing as true seems prone to inter-subject variability (I find (16) rather dubious; Lewis obviously does not). However, Conditional Excluded Middle is a direct consequence of the Uniqueness Assumption, and Lewis’ attack on the latter is more convincing. Consider (17):

(17) If I were taller, I would be good at basketball.

Is there a unique possible world in which I am taller holds, and which is more similar to this world than any other world is? There cannot be; for if I am x inches taller at world α than I am in the actual world, there will exist a near-identical world β in which I am only \( \frac{x}{2} \) inches taller than I am in the actual world. There is an infinite number of ever-closer possible worlds, but no one closest world. Nor can there be a set of equally close worlds that are closer than any other worlds.

Lewis’s (1973) analysis is directly motivated by the need to account for examples such as (16) and (17). His semantics has much in common with Stalnaker’s; both posit a set of possible worlds that are ordered with regard to similarity. Though Lewis breaks with convention and does not directly allude to the heritage of the Ramsey Test, a spirit of continuity with previous accounts is evident in his opening statement:

‘If kangaroos had no tails, they would topple over’ seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. (Lewis 1973: 1)

A Lewis model for counterfactuals is a triple \( < I, $, [], > \), where

- \( I \) is the set of all possible worlds.
- \$ is a centered system of spheres which assigns to each \( i \in I \) a set \( S_i \) of sets of possible worlds (spheres) that are centered on and nested around \( i \), i.e. \( i \in S_i \) and \( \forall S, T \in S_i [ S \neq T \rightarrow ( S \subset T \lor T \subset S ) ] \), respectively.
- \([ \] \) is an interpretation as in Stalnaker’s system.

Lewis introduces two counterfactual connectives: \( \Box \rightarrow \) and \( \diamond \rightarrow \). The former corresponds to a would-counterfactual (If \( \phi \) were the case, then \( \psi \)
**CHAPTER 2. PHILOSOPHICAL PERSPECTIVES**

would be the case), the latter to a might-counterfactual (If \( \phi \) were the case, then \( \psi \) might be the case). Functionally, they are analogous to conditional necessity and possibility in a given sphere. I shall concentrate on \( \Box \rightarrow \), whose truth conditions Lewis defines as follows\(^{10}\) (p. 16):

\[
\phi \Box \rightarrow \psi \text{ is true at a world } i \text{ (according to a system of spheres } S) \\
\text{if and only if either} \\
\text{(1) no } \phi \text{-world belongs to any sphere } S \text{ in } S_i, \text{ or} \\
\text{(2) some sphere } S \text{ in } S_i \text{ does contain at least one } \phi \text{-world, and} \\
\phi \rightarrow \psi \text{ holds at every world in } S
\]

In order to evaluate a counterfactual, it is no longer sufficient to investigate a single antecedent-world (as we have abandoned the Uniqueness Assumption), nor may we assume a finite set of maximally similar worlds. Lewis wields his similarity metric in an outward direction; the concept of a sphere represents the imposition of an outer boundary, inside which every world meets some criterion of overall similarity to the actual or base world \( i \). The larger the sphere, the less strict the similarity criterion, and vice versa. A counterfactual is non-vacuously true if a similarity criterion can be found whereby the corresponding material conditional holds in every world that meets that criterion. If the antecedent is true in the actual world, then the counterfactual is true iff the consequent is true in the actual world, and it reduces to the material conditional. A counterfactual is vacuously true iff the antecedent is not true in any accessible possible world, i.e. it is impossible (the nature of the ‘impossibility’ will be determined by the nature of the accessibility relation).

Lewis describes a set of ‘\( V \)-logics’, within which a formal semantics for his counterfactual operators and other modal machinery may be given. His logic of counterfactuals is the system \( VC \). Modus Ponens is valid in \( VC \), as is Modus Tollens; Contraposition, Antecedent Strengthening and the Hypothetical Syllogism are not. As previously noted, Conditional Excluded Middle is not valid in \( VC \).

Again, the bulk of the explanatory work is left implicit; it is just assumed that the system of spheres implementing the similarity ordering on possible worlds is already in place. Lewis (1986b) maintains that the analysis in Lewis (1973) is ‘about all that can be said in full generality about counterfactuals...it does little to predict the truth values of particular counterfactuals in

\(^{10}\) \( \phi \Diamond \rightarrow \psi \) is equivalent to \( \neg (\phi \Box \rightarrow \neg \psi) \).
particular contexts... It must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context” (p. 41). Lewis (1986a) makes the vague remark that “we may say that one world is closer to actuality than another if the first resembles our actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off against one another” (p. 163). Furthermore, some aspects of the world are more important than others; Lewis (1986b) claims that the vowel/consonant ratio counts for nothing when comparing the writings of Wittgenstein and Heidegger\textsuperscript{11}. Sanford (1989) expresses scepticism about the entire possible worlds endeavour, and dwells on the question of similarity (p. 151-2):

Are there some objective facts about possible worlds and comparative similarity between worlds that support a definite answer although they happen to elude our intellectual grasp? A bold metaphysician who says so is too bold to be believed.

Lewis (1986b) attempts to sketch some criteria for resolving the vagueness inherent in allusions to similarity. He posits a ‘standard’ mode of similarity resolution, which may be overridden by ‘special’ resolutions should the discourse context demand it. By analysing how he might wish certain counterfactuals to be interpreted, he arrives at the following directions for weighting worlds with regard to similarity:

1. It is of the first importance to avoid big, widespread, diverse violations of law.

2. It is of the second importance to maximise the spatio-temporal region throughout which perfect match of particular fact prevails.

3. It is of the third importance to avoid even small, localised, simple violations of law.

4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

(Lewis 1986b: 47–8)

\textsuperscript{11}Though one can readily conceive of conversational contexts where this difference would be most salient; one must assume that such contexts are ‘special’ or marked.
Two matters need elucidation before we can consider the implications of this proposal. Firstly, a satisfactory definition of lawhood is required. Lewis (1973: 73) proposes that “a generalization is a law at a world \(i\) . . . if and only if it appears as a theorem in each of the best deductive systems true at \(i\).” This raises further questions: what is the ‘best’ deductive system that is true at a world? In Lewis’ view, it is that system which has the optimum balance of ‘simplicity’ (generality) and ‘strength’ (informativeness), and which allows all true facts of that world to be deduced. We cannot know such laws (lacking omniscience); however, we can and do assume that the events of our world are governed by some set of ‘laws of nature’. Furthermore, we can intuitively accept that a world whose worlds differ significantly from ours will be significantly dissimilar to ours under any normal conception of similarity (cf. Lewis’ weighting criterion 1)\(^{12}\). If (as Lewis does) we assume a deterministic universe, the upshot of this conception of lawhood is that whenever two worlds with completely identical histories diverge, we must conclude that the worlds are subject to different laws.

Secondly, we come to the nature of causation. It is unclear how Lewis’ four criteria can help resolve counterfactuals like the Apollo-sentences in (15) without recourse to omniscience. The main purpose of his system for world-ordering is to explain counterfactuals that directly involve causation. Lewis (1986a) presents an analysis of causation that explicitly invokes counterfactuals\(^{13}\), and the ordering in Lewis (1986b) reflects this analysis, and his desire to defend the asymmetry of causation: a cause may have many effects, but an effect may have only one cause. This is equivalent to saying that the past is fixed, or overdetermined by the present.

Lewis (1986b) discusses the counterfactual in (18), due to Kit Fine:

\(^{12}\)When Lewis discusses ‘miracles’ or ‘law violations’, he means that the laws of the counterfactual world diverge from those of the actual world in some manner. Hence the laws of the actual world are seen to be violated at another world. It is clearly impossible for the laws of a world to be violated at the same world.

\(^{13}\)Lewis’ account is as follows: we say that an event \(e\) causes an event \(c\) iff the following counterfactuals are true:

1. \(O(c) \not\rightarrow O(e)\)

2. \(\neg O(c) \not\rightarrow \neg O(e)\)

where \(O(e)\) is a proposition that holds at all and only those worlds where the event \(e\) occurs.
(18) If Nixon had pressed the button there would have been a nuclear holocaust.

An apparent problem for Lewis arises here because one might hold that a possible world in which Nixon presses the button, but where a nuclear holocaust fails to transpire, would be closer to the actual world than one in which there was a nuclear holocaust (nuclear holocausts presumably leading to significant differences between worlds). However, it is just that holocaust-world that we need to validate the counterfactual. Lewis examines a number of candidate antecedent-worlds with regard to which (18) might be evaluated, and thus arrives at his four criteria. His ‘most similar’ world is that which proceeds exactly as ours until a timepoint $t_1$. At $t_1$ a small, localised ‘miracle’ occurs, which causes Nixon to press the button, and a nuclear holocaust ensues. The laws of this world ($w_1$) are exactly the same as those of the actual world, with the sole exception that they license the miracle at $t_1$. If the deterministic laws of the counterfactual were wholly identical to those of the actual world, Nixon could not press the button unless the entire past and future of the worlds were at every timepoint divergent$^{14}$. Lewis also considers a world $w_3$ where the miracle once again occurs at $t_1$, but where a second miracle occurs at a subsequent time $t_2$, removing all traces of the first miracle, and causing the counterfactual world to converge to the actual, so no holocaust occurs. Lewis must make a distinction between ‘small’ miracles of divergence and ‘large’ miracles of convergence, and asymmetry of causation offers him a way. A miracle of convergence is ‘large’ because it must undo every effect of the divergence miracle (which will be many); a miracle of divergence need only alter the world in one place$^{15}$.

$^{14}$Although Lewis’ analysis rests on the assumption of determinism, he claims that it would also hold in an indeterministic universe (Lewis 1986b, Postscript C).

$^{15}$Lewis’s (1986b) examples of ‘small’ miracles are to my mind misleading. He imagines (p. 44) that “a few extra neurons fire in some corner of Nixon’s brain”, causing him to press the button. But multiple neural firing is by no means a simple event, and may involve numerous heterogeneous events, and seems to fit his description of a ‘big miracle’: “What makes the big miracle more of a miracle is not that it breaks more laws; but that it is divisible into many and varied parts, any one of which is on a par with the little miracle” (p. 56). Lewis can easily rescue the situation, however, if he suggests that a previous, sufficiently small miracle ultimately causes the relevant neurons to fire in opposition to their non-firing in the actual world. This may lead us to the question: how small is small? It is tempting to assert that small miracles must occur on the atomic level (or maybe lower?); however, what is crucial for Lewis is that small miracles, on some sufficiently low level, happen just once and just in one place.
Lewis has (to my knowledge) been the only philosopher brave enough to stick his head above ground and offer a method for judging similarity. Yet that has not prevented others raising their objections. Krasner & Heller (1994) present a number of counterfactual sentences, which they claim are counterexamples to Lewis’ world-ordering. For example (their Counterexample C, p. 36):

(19) (spoken after a nuclear holocaust, presumably at the world \( w_1 \) from (18)) If there were as many people today as there were back in 1960, there would not be enough food or water for them, and the vast majority of them would be living short and miserable lives.

Krasner & Heller point out that the easiest way of satisfying the antecedent of (19), under Lewis’ weightings, would be to have a small miracle prevent Nixon pressing the button and causing the holocaust, thus ensuring perfect spatio-temporal match up to the timepoint when the miracle would occur\(^{16}\). Yet we do not wish to evaluate the counterfactual with regard to a world where no holocaust occurs. Hence, Lewis’ criteria seem to select the wrong world. This example is reminiscent of Chisholm’s Apollo-sentences, in that the similarity metric required is strongly influenced by our wish to satisfy the consequent. Lewis might call this a situation where ‘special’ vagueness resolution is required, and hope that this explanation suffices. However, (19) surely undermines Lewis’ criteria, if any case they do not account for is deemed to require a ‘special’ kind of similarity. Counterfactuals are so deeply context-dependent that Lewis’ system, providing a weighting for counterfactuals independent of context, may ultimately be of little use. If counterexamples cannot validly contradict the system, then the system must stand or fall on Lewis’ account of causality.

A number of philosophers have signalled unease with Lewis’ adherence to the asymmetry of causation. Lewis claims that ‘backtracking counterfactuals’ (involving backward causation) are rarely, if ever, acceptable. Yet counterfactuals like (20) seem valid:

(20) If the lamp had come on, I would have had plugged it in earlier.

(Sanford 1989:183)

Sanford accordingly asserts that as at least some backtracking counterfactuals are permissible, “the asymmetry of overdetermination is an illusion”\(^{16}\)Unless, of course, a suitably small miracle could cause millions of people to appear at some more recent timepoint.
I would suggest that the situation is not so clearcut as Sanford might wish. While (20) is certainly acceptable, it is acceptable in virtue of the fact that there is only one reasonable way for a lamp to come on. In this case, there is little overdetermination of the past by the future, but this case is not representative. The world is a bafflingly complex system, and one way in which individuals simplify their conceptions is by thinking of the past as 'fixed' and the future as 'mutable'; this may not satisfy a metaphysician, but it may satisfy a cognitive scientist. Ultimately, the question of backtracking resolves to the question of 'necessary connections'. Sanford posits a reflexive pattern of necessary dependence between the lamp coming on and its being plugged in; Lewis must say that the world most similar to ours in which the light comes on is one in which it is plugged in. After all, backtracking conditionals are only ever facilitated by 'special' vagueness resolution (Lewis 1986b: 34).

Nute (1984) describes a backtracking scenario that causes further problems for Lewis: I left my coat unattended in a certain room yesterday. Today I return and find that it is still there. Lewis' world ordering will now dictate that had it been taken, it would have been taken today rather than yesterday. Obviously, there is no reason for this to be true. For example, it may be the case that I know many people were in that room yesterday, and none were there today. Lewis can respond that this is clearly a 'special' situation, firstly as his criteria were never intended to handle backtracking counterfactuals, and secondly contextual information is impinging on my similarity judgement (but as no real-life counterfactual is immune from prior knowledge, this second line of defence is unattractive).

To sum up: I have devoted a great amount of space (and will devote more) to the shortcomings of Lewis’ account. This may not be wholly fair, as it remains the most explicit (and most influential) account of counterfactuals yet proposed. I will return to it in later sections.

2.4 Partial Representations

When we entertain a counterfactual state of affairs, only some aspects of the actual world seem relevant. For example, if I think about what would have happened if I had remembered to take my umbrella when I went out, I am unlikely to consider the weather in Brazil, or the market price of crude oil, or how many leaves are on the tree across the road. In discourse, there is
usually an implicitly demarcated context, within which definite descriptions and (sometimes) ambiguous phrases have a commonly understood meaning; it is not permissible to refer outside the discourse context without explicitly signalling that the context is being expanded. Likewise with counterfactuals. Some authors have advocated a semantics for counterfactuals based on partial representations which, unlike possible worlds, are not maximally specified.

Sanford (1989) adopts a position of ‘modal super-realism’ (in contrast to David Lewis’ modal realism). This involves positing a multitude of spatio-temporally connected areas called WORLDS, all of which are contained in the actual world. Accordingly, a conditional like If I had dropped the egg, it would have broken “does not need a world thousands of light-years across and millions of years old. A few minutes in a WORLD the size of a kitchen would be enough” (p. 165). Recourse to additional possible worlds is unnecessary; the entities referred to in any counterfactual will already be present (at least conceptually) in this world. Sanford describes his approach as a ‘one-world theory’. He claims that WORLDS are non-overlapping, yet it is hard to see how this assertion is tenable, if we are to assume that counterfactuals about, say, the European continent and the Russian state should be evaluated in different WORLDS. In any case, the concept of partial WORLDS is not integral to the substance of Sanford’s theory of conditionals, which holds that a conditional (be it factual or counterfactual) is true if a pattern of dependence between the antecedent and consequent can be satisfied. For example, five distinct patterns of dependence can be identified in the conditional If Smith was fired, so was Wilson:

1. Smith’s firing is a sufficient condition for Wilson’s firing.
2. Wilson’s firing is a necessary condition for Smith’s firing.
3. Smith’s firing is a sufficient condition for Wilson’s firing and Wilson’s firing is a sufficient condition for Smith’s firing.
4. Some necessary condition for Smith’s firing was also a sufficient condition for Wilson’s firing.
5. Independent causation; If anyone was fired, Wilson was fired.

The conditional is true if any one of these patterns holds; the nature of the pattern that validates it plays no role in determining its truth or falsity.
Just which pattern is identified in a conditional will be mediated by world knowledge and other pragmatic considerations; Sanford does not specify how this is done. Although Sanford disagrees with Lewis on the acceptability of backtracking dependences\(^{17}\), patterns of dependence could emerge from a suitable similarity metric over possible worlds; however, this would obviously require an ontological shift which Sanford would presumably find unacceptable. A metalinguistic explanation might be more appealing, though this would require a definition of cotenability other than that to be described in section 2.5, which depends on possible worlds. Sanford’s theory also faces the problem which haunts all ‘necessary connection’ theories, that of explaining the truth of conditionals like (10), which are true solely in virtue of their true antecedents. One possible solution to this problem would be to abandon the conventional meaning of ‘dependence’ and adopt a definition similar to that of ‘conditional constraints’ in situation semantics (see below). If we do this, then Sanford’s theory essentially becomes a less developed version of that account (WORLDS could easily be formalised as situations).

A promising approach to counterfactuals and other conditionals is Barwise’s (1989) situation semantics account. The primary focus of research in situation semantics is the information content of utterances, broadly understood as what a given utterance communicates about the context in which it is uttered. Information is therefore an epistemic concept. The information content of an utterance can vary from context to context, and may vary with individuals’ perspectives within a single context. This contrasts with the absolute truth conditions investigated by classical semantics. As the name suggests, situation semantics deals with situations, “those portions of reality that agents find themselves in, and about which they exchange information.” (Barwise 1989: xiv). It follows from this definition that situations are partial with respect to the actual world, the scope of a situation being determined by the discourse context.

Situation semantics gives an explanation of meaning as follows (p. 81): the interpretation of a sentence \(S\) is the situation type \(T_S\) described by \(S\). For example, the interpretation of \(John\) \(walks\) is the type of situation in which John walks. To assert \(S\) with an utterance \(u\) is to assert that the situation \(s_u\) with regard to which the utterance is made is of the situation type \(T_S\). The information content of \(u\) is that \(s_u\) is of type \(T_S\); \(u\) is true iff \(s_u\) is of

\(^{17}\)Explicitly so: “We should regard some of them [backtracking conditionals] as true or as acceptable as any forward-tracking conditional” (p. 187).
CHAPTER 2. PHILOSOPHICAL PERSPECTIVES

29

type $T_s$. The role played by the utterance situation $s_u$ reflects the degree to which contextual factors (e.g. indexicals, anaphora) influence the way an utterance is understood.

Barwise (1989: ch. 5) presents a unified semantics for conditional sentences. Conditionals are held to describe a constraint between situation types – *if* $p$, *then* $q$ means that every situation of the type described by $p$ is also of the type described by $q$. More formally, the constraint $T_1 \Rightarrow T_2$ (where $T_1$ and $T_2$ are situation types) means that $T_1$ involves $T_2$, in keeping with the Principle of Involvement: if $s$ is a situation of type $T_1$ and $T_1$ involves $T_2$ then there is a compatible situation of type $T_2$ (Seligman & Moss 1997: 300). This principle rests on the concept of compatibility; two situations are said to be compatible if they are part of some larger situation. Seligman & Moss observe that the interpretation of this concept is still controversial in situation theory. Under one interpretation, compatibility is the same as compossibility, and two situations are compatible if they are both part of some maximal situation. A ‘maximal situation’ is essentially a possible world, and compossibility becomes analogous to cotenability at a single world. Under a different interpretation of compatibility, incompatibility arises as a result of differing perspectives on the same world, and it may be possible to reconcile the incompatible situations by taking account of the differences in perspective.

The meaning of (21a) is the constraint specified in (21b); to assert (21a) is to assert that the constraint is actual in a given situation. The parameter $l$ in (21b) takes some set of locational values within the location specified or otherwise indicated by the speaker; the 1 represents truth, i.e. that the given situation type obtains.

(21)  a. If it snows, then the sidewalks are slippery. (Barwise 1989: 123)
          b. $C = (S' \Rightarrow S'')$
             $S' = [s|in \ s: at \ l: snowing;1]$  
             $S'' = [s|in \ s: at \ l: slippery, sidewalks;1]$

The contextual vagueness that haunts all accounts of conditionals is manifested as a situation type $B$ that corresponds to a set of background conditions determined (‘anchored’) by the environment against which a conditional is uttered. So long as the referent situation is of type $B$, then the conditional constraint is actual. Background conditions are not explicitly described by the speaker, and the hearer must decide which background conditions are
appropriate. For many indicative conditionals this will not be difficult, as
the utterance will usually be about the current situation, and the properties
of that situation should be perceived similarly by both parties. The ease
with which assertions about non-immediate situations will be felicitously in-
terpreted will depend on the degree of common beliefs held by speaker and
hearer about the situation in question. Counterfactuals provide yet more
of a problem, as the hearer is faced with the need to ascertain just which
background conditions the speaker has in mind such that the conditional
constraint \( C|B \) is actual, and to ascertain what counterfactual situation (of
type \( B \)) the utterance is about. We can now explain the Apollo-sentences in
(15): both sentences are true with regard to different background conditions,
specifying different criteria for manhood.

Barwise assumes that even when we speak counterfactually, we are still
discussing relations between entities in the actual world, and that there
is no need to commit to the existence of other possible worlds\textsuperscript{18}. Barwise
(1989: 82). Yet it is not made clear how we are to investigate whether, for a
counterfactual conditional constraint \( S' \Rightarrow S''|B \), every situation of type \( S' \)
will also be of type \( S'' \). We may have access to other actual situations of type
\( S' \), but we may just as soon not have access to such situations. No situation
of the type \([s|in s: in 1937: at l: the CIA assassinate Hitler;1]\) has ever been
realised in this world (cf. (3)). If we are to reason logically about what
follows from the facts of situations of type \( S' \) together with the background
conditions \( B \), then we are essentially committing ourselves to a metalingu-
istic solution. Another way out is to adopt a purely epistemic approach, and
to locate situations in the perceptual world, not excluding our perceptions
of illusions and imaginary states of affairs. I find some tentative support for
this direction in Cavedon (1995): “Unlike a possible-world, a situation does
not describe the way the world could be, but the way the world is, or at least
the way it is perceived to be” (p. 4; last italics added by this author).

\textsuperscript{18}The compossibility interpretation of compatibility lends itself to a possible worlds
framework. Kratzer (2002) describes a ‘premise semantics’ which posits situations as well
as possible worlds. Her treatment of counterfactuals is, however, most reminiscent of
metalinguistic approaches: “In a premise semantics, a ‘would’-counterfactual is true in a
world \( w \) iff every way of adding as many facts of \( w \) to the antecedent as consistency allows
reaches a point where the resulting set logically implies the consequent.” Her analysis of
facthood rests on a distinction between particulars (situations) and propositions (which
seem to resemble situation types); ‘facts’ are held to be propositions, not particulars. This
is not ultimately dissimilar to Barwise’s view.
Seligman & Moss (1997) are sceptical about the tenability of the conditional constraint analysis. They note the difficulty involved in identifying the background conditions that obtain in a given situation. Furthermore, they note the difficulty in separating the information that a situation is of type $B$ from the information that the antecedent and consequent are actual.

Nevertheless, Barwise’s proposal has many merits, not least its similarity to one of the psychological theories considered in chapter 3. It offers a unified semantics for both factual and counterfactual conditionals. It explains how classically valid inference patterns involving conditionals sometimes fail, and sometimes do not, in terms of ‘background conditions’ (see section 2.5).

2.5 Discussion

The theories described in the previous sections all have much in common. This should not be surprising, as any attempt to resolve the problem of counterfactuals will have to confront the same issues. Three such issues seem especially salient: (i) the relationship between factual and counterfactual conditionals; (ii) the vagueness inherent in the interpretation of counterfactuals; and (iii) the validity of conditional inference rules.

2.5.1 Factual and Counterfactual Conditionals

Factual and counterfactual sentences constitute distinct but not wholly dissimilar categories. Both are kinds of conditionals; both signal invitations to hypothetical thought. Yet we can always decide the truth or falsity of a factual conditional by consulting the actual world, or at least with our beliefs about the actual world. Whether a counterfactual can be so decided is contentious; the answer depends on the nature of the representation one favours for the logical structure of the world and our beliefs about it. Strawson (1986: 230) considers the two sentences in (22), and observes: “It seems obvious that about the least attractive thing that one could say about the difference between these two remarks is that it shows that, or even that it is partly accounted for by the fact that, the expression ’if... then’ has a different meaning in one remark from the meaning which it has in the other.”

(22) a. Remark made in the summer of 1964: ‘If Goldwater is elected, then the liberals will be dismayed.’
b. Remark made in the winter of 1964: ‘If Goldwater had been elected, then the liberals would have been dismayed’

The argument runs as follows: the meaning of the conditionals does not change; rather, the facts of the actual world have changed between the time that (22a) is uttered and the time that (22b) is uttered. What was still a factual possibility at the time of (22a) has become a counterfactual possibility by the time of (22b), but the facts that allowed us evaluate (22a) should still be available to allow us evaluate (22b). Therefore, we should favour a unified theory of conditionals. If possible worlds are used to evaluate one conditional, they should be used to evaluate the other; if they are unnecessary for one, they should not be necessary for the other.

Other philosophers have argued that counterfactuals are not always equivalent to their factual counterparts. A much-cited counterexample (due to Adams (1970)) is (23); sentence (23a) seems true, whereas sentence (23b) may not be.

(23)  
   a. If Oswald did not kill Kennedy, then someone else did.
   b. If Oswald had not killed Kennedy, then someone else would have.

Lewis (1973: 3) finds this example sufficiently convincing to rule out any unified analysis of conditionals. Barwise (1989: 107) disagrees, on the basis that (23a) is not the proper indicative conditional with which (23b) should be compared; rather, (23a') is:

(23a') If Oswald has not killed Kennedy, then someone else will have.

(23a') is to my mind a rather odd-sounding sentence. With the exception of temporality, it does not seem to differ in any interesting way from (23a), and with respect to temporality, it seems less pleasing as a counterpart to (23b) than does (23a). It is doubtful that anyone today would choose to assert (23a')\(^{19}\), whereas both (23a) and (23b) are perfectly acceptable. If we adopt the classical Reichenbachian treatment of tense, then (23a) and (23b) assert the same temporal relations\(^ {20}\); (23a') asserts a different temporal relation.

---

\(^{19}\)I am not aware of the standard treatment of dead (or as yet unborn) people in formal semantics. However, it does not seem permissible to assert that an individual plays a role in an event that occurs for a duration including time \(t\), when that individual is not alive at time \(t\).

\(^{20}\)I am accepting the analysis of the ‘counterfactual pluperfect’ in Steedman (1997). Like the past tense, the counterfactual pluperfect posits coinciding reference and event
What, then, underlies the difference in acceptability between (23a) and (23b)? I would contend that the only difference is that (23b) presupposes in some way that Oswald killed Kennedy, and that (23a) does not. To assert (23a) is to assert, on the basis of relevant knowledge and beliefs (presumably including the knowledge that Kennedy was killed) and a way of measuring similarity between worlds, that for some set of privileged possible worlds, in every world where Oswald did not kill Kennedy, someone else did; or, that the antecedent, conjoined with relevant knowledge, entails the consequent; or, that a conditional constraint holds such that every situation in which no temporally preceding situation was one where Oswald kills Kennedy is a situation in which someone else did. To assert (23b) is to assert, on the basis of relevant knowledge and beliefs (presumably including the strong belief that Kennedy was bound to be killed at some point) and a similarity metric, that for some set of privileged possible worlds, in every world where Oswald did not kill Kennedy, someone else did; or, that the antecedent, conjoined with some cotenable relevant knowledge and beliefs, entails the consequent; or, that a conditional constraint holds such that every situation in which no temporally preceding situation was one where Oswald kills Kennedy is a situation in which someone else did.

I realise that this interpretation of possible worlds semantics differs significantly from Lewis’. It is in some ways closer to Stalnaker’s theory, in that it assumes a unified theory of conditionals. This does not prevent us endorsing Lewis’ rejection of the Uniqueness Assumption and Conditional Excluded Middle. It does commit us to rejecting the material conditional interpretation of indicative conditionals. For example, only necessarily false antecedents make a factual conditional vacuously true. Factual conditionals with antecedents that are contingently false in the actual world may be true or false. A sentence like \textit{If the CIA killed Hitler in 1937, the world is round} is presumably true; \textit{If the CIA killed Hitler in 1937, the world is flat} is false; \textit{If the CIA killed Hitler in 1937, Hitler was dead by the summer of 1940} is true. In effect, factual conditionals with false antecedents become equivalent to the corresponding counterfactuals; this parallels Lewis’s (1973: sec. 1.7) proposal that apparent counterfactuals with true antecedents be evaluated on the same basis as the corresponding factual conditionals. I concede that indicative conditionals with clearly false antecedents appear odd; but this

\footnote{timepoints that precede the time of utterance; the present perfect (as in (23a')) posits coinciding reference and utterance timepoints that are preceded by the time of the event.}
can be related to the associated conversational implicature to the effect that neither the falsity of the antecedent nor the truth of the consequent are known.

The elimination of the factual-counterfactual distinction brings the possible worlds analysis into line with the metalinguistic theory. It was noted that the main problem facing the metalinguistic account was seen to be the lack of an adequate definition of cotenability. Lewis (1973: 57) shows that cotenability can be defined in terms of his possible-worlds system: “Let us say that $\chi$ is cotenable with $\phi$ at a world $i$ (according to a system of spheres $\$)$ if and only if either (1) $\chi$ holds throughout $\bigcup \$i$, or (2) $\chi$ holds throughout some $\phi$-permitting sphere in $\$i$.” It follows that the metalinguistic analysis (given this version of cotenability) and Lewis’ analysis are functionally equivalent. Lewis defends his approach on grounds of parsimony. In the metalinguistic case, one must first tackle the vagueness inherent in underspecified antecedents, and then resolve the issue of entailment; in the Lewisian case, these two problems reduce to one, the problem of inter-world similarity. Lewis concludes: “Better one sort of influence of context than two different sorts” (p. 67). Whether one accepts this judgement or not, it is clearly far from a knock-down argument. The situation semantics equivalent of cotenability is compatibility, the interpretation of which is, as we saw earlier, not universally agreed on. Questions of compatibility are not resolved by investigating other possible worlds, but by seeing whether two situations are part of some larger situation, which may or may not be a world. It is approvingly noted that situation semantics also offers a unified theory of conditionals.

### 2.5.2 Conversational Vagueness

The problem of vagueness in counterfactuals reflects the general problem of vagueness in language. This is perhaps the fundamental problem of symbolic communication: how can a speaker encode the information he wishes to convey as effectively and as efficiently as possible? The speaker, having the luxury of total access to his thoughts, will wish to keep his utterances as compact as he can; he will see no need to offer information that he sees as tangential or peripheral. His privileged mental access (and lack of access to the hearer’s stock of beliefs) will in many cases lead him to underestimate the amount of information that is needed to help the hearer resolve the vagueness in his utterance. In Gricean terms, there is a tension between the second Maxim of Quality (“Do not make your contribution any more informative
than necessary”) and the first Maxim of Quality (“Make your contribution to the conversation as informative as is required”) together with the Maxim of Manner’s exhortation to “Avoid ambiguity”.

Vagueness is rife in natural language\textsuperscript{21}. The sentence *John saw the girl with the telescope* does not by itself tell us who had the telescope. Lexical items may refer to cluster concepts or prototypes, such that the criteria that permit the use of an item in one case may not be the same as permit the use of that item in another case (the classic example is Wittgenstein’s (1953) discussion of the meaning of the term *game*\textsuperscript{22}). This leads to vagueness from the hearer’s perspective – if I am told that *x* is a game, what can I conclude about *x*? There is also vagueness from the speaker’s perspective, where it may not be clear which term should be used to refer to an object – when does a game become a sport? In many of these cases, different people will have different criteria for determining the boundaries of a lexical item’s meaning. Sometimes (as in the distinction between fruit and vegetable) there is a normative criterion; yet this criterion may not always be accepted. I may or may not know that a cucumber is technically a fruit, but whether I do or not, it might just not feel right to assert this. Definitional vagueness is by no means restricted to nouns: for example, Fillmore (1982) gives a prototype semantics for the verb *climb*. Some terms are inherently vague, and can only be applied relative to some contextually mediated metric. Lewis (1979) recounts Unger’s (1975) contention that almost nothing is truly flat – no matter what entity we might wish to describe as ‘flat’, some other thing is bound to be flatter (within physical limitations), so how can we acceptably insist on the flatness of the entity?

We must conclude that any linguistic symbol or string is employed with

\textsuperscript{21}I recognise that I am departing from the technical usage of the word *vague*, which as I understand it refers to properties whose borders are not determinate. I am willing to conflate this concept with that of ambiguity (which refers to symbols or strings whose meaning is not determinate), for purposes of convenience. Within the strict confines of this discussion, I do not find this abuse of terminology to be overly objectionable. Both sorts of indeterminacy tend to be resolved via the same kind of contextual inferences.

\textsuperscript{22}Wittgenstein points out that the word *game* does not really have a definition; rather, we judge that something is a game on the basis of its ‘family resemblance’ to other games. No set of necessary and sufficient conditions that we can think of will be general enough: “Look at board games, with their manifold similarities. Now proceed to card games: here you find many correlations with the first class, but many common properties disappear, and others appear. If we move on to ball games, we retain some commonalities, but many are lost” (§66; my translation).
a degree of vagueness. Whether or not its use is conversationally acceptable will depend on whether the entity or state of affairs it is being used to refer to sufficiently resembles its ‘core meaning’, as given by the internalised grammar (including the lexicon) of each party in the conversation. If we accept Quine’s (1960) arguments for the indeterminacy of translation, then this systemic vagueness becomes advantageous, indeed necessary, for linguistic communication. If two individuals can never be sure that a given word or string has the same meaning for both of them, then they will not be able to communicate unless the interpretation of the word or string is flexible. So long as they do not encounter a situation where their interpretations become overly divergent\textsuperscript{23}, none of this vagueness is even noticed.

Vagueness may be necessary for language, but the ability to resolve vagueness is just as necessary. Luckily, people are generally adept at dealing with vagueness and ambiguity. Our unconscious ability to resolve anaphora and definite descriptions far surpasses the conscious attempts of linguists to make this ability formally explicit. We can agree that a table is flat even when the table isn’t actually flat, and when our internal representations of flatness differ. However, things are sometimes just too unclear, e.g. Seamus and Joe were standing on the street, and then he hit him. Counterfactual conditionals frequently test the limits of our resolution abilities.

A crucial role in resolving vagueness is played by the speaker’s and hearer’s mutual knowledge and beliefs, i.e. their common ground. One component of this set will be the perceptually available facts about the immediate environment (assuming that both parties are in the same place, and not communicating via telephone, email, letter, etc.). Given that the hearer accepts the reliability of the speaker and possesses a sufficient degree of charity towards him, he will accept an assertion that does not contradict his beliefs. If an assertion does contradict some of the hearer’s beliefs (which will presumably not be part of the common ground), the hearer must decide to deny the assertion and explain his contradictory beliefs, or to accept the assertion and revise his stock of beliefs, or to accept the assertion on a hypothetical basis for the sake of the argument. Which decision he reaches will depend on a number of factors, including the comparative reliability of the speaker and the source of the contradictory prior belief, and the comparative cotenability

\textsuperscript{23}Unfortunately, there are situations in which conversational flexibility can have disastrous consequences; Cushing (1994) describes how the use of vague or ambiguous language by aircraft pilots and air traffic controllers can (and has) lead to tragic accidents.
of the new assertion and the prior belief with the rest of the hearer’s beliefs. Once an assertion made by one party has been accepted by the other(s), it is added to the common ground. This is simply how normal conversations work; so far as I am aware, it is not controversial in the literature.

Counterfactuals are more prone to problems of vagueness than are other conditionals because they refer to situations which contradict the speaker’s beliefs about the world. Of course, they will usually contradict the hearer’s beliefs as well, but occasionally they may not, in which case the conversation must first establish which version of events is going to be accepted before the counterfactual can be discussed. Evaluating a counterfactual thus involves revising one’s beliefs to facilitate the antecedent. In scenarios triggered by conditionals like If I had pressed the light switch, the light would have come on, a single way of revising beliefs may be particularly obvious. In others, like the Apollo sentences, there may be more than one equally natural way of doing this. When confronted with the assertion that had Caesar been in command in Korea, he would have used catapults (Quine 1960), the hearer’s task is to decode both the meaning of the sentence (which should not be difficult) and the information (similarity metric, set of relevant beliefs, conditional constraint) that licenses the conditional. Our present concern is with the latter part. By the Cooperative Principle (Grice 1975), the hearer should do his/her best to find a similarity metric, etc., that renders the counterfactual true. Though he will be charitable, the hearer should not be willing to abandon too many well-founded beliefs in order to force the counterfactual into an acceptability-shaped hole – it would be wrong to abandon the belief that catapults were not used in the Korean war. Being polite, the cooperative hearer who cannot bring himself to verify an asserted counterfactual will recognise the vagueness that is involved and will not immediately assert its falsity; rather, he should ask How does that follow?, or What exactly are you assuming?²⁴

This process of identifying a satisfactory similarity metric, etc., will clearly be guided by the consequent of the counterfactual. To evaluate (15a), we will seek a metric that makes Apollo mortal; to evaluate (15b), we will seek a metric that makes Apollo immortal. As was just noted, this cannot be the

²⁴It would seem odd to ask How exactly are you evaluating similarity between worlds? That we explicitly ask for assumptions rather than similarity metrics when resolving vagueness may circumstantially speak for the metalinguistic theory against a possible worlds theory. But as we have seen, a metalinguistic theory that does not refer to possible worlds faces significant problems.
only criterion that guides our search; otherwise, we would happily assent to all counterfactuals. Yet it must be an important factor. It seems reasonable to suggest that this consideration might interact with other criteria (such as those of Lewis 1986b) to determine a ‘biased’ similarity metric. This does not go far towards a satisfactory account of how similarity (etc.) is decided – this will ultimately require an account of how facts and beliefs are weighted – but it might help to explain the acceptability of (17), repeated here:

(17) If I were taller, I would be good at basketball.

Without the antecedent-satisfaction criterion, it is hard to see how any standard similarity metric could render this sentence true. Assuming that I need to be 2cm taller than I am now to be good at basketball\textsuperscript{25}, merely being 1cm taller would not make me good. Yet a world in which I am 1cm taller should be more similar to the actual world than one in which I am 2cm taller. My antecedent-satisfaction criterion gives us a more satisfactory way to establish similarity. One might object that (17) is not really true, given these considerations, but it seems no more vague or delinquent than most other counterfactuals. This conversationally-driven consideration may be identified with Lewis’s (1986b) ‘special’ mode of vagueness resolution, or with one element thereof.

In the latter part of the above discussion, I have argued with reference to possible worlds, mainly because of their explanatory advantages. This does not mean that my arguments do not apply to the metalinguistic or situation semantic accounts; on the contrary, they are just as salient. Deciding how to expand an antecedent or identifying background conditions is equivalent to establishing a similarity ordering. If I decide that my belief that the earth is round is relevant to a counterfactual, and that I am unwilling to abandon that belief under any circumstances, then I will judge round-earth possible worlds to be more similar to the actual world than square- or dodecahedral-earth worlds are. If, as in (17), the shape of the earth does not seem relevant, this belief of mine will not play any significant role in distinguishing between worlds.

I surmise that many of my views, and this last point in particular, will have appalled any metaphysician who has read this far. Surely a difference in the shape of the earth, and the physical consequences of that difference,

\textsuperscript{25}The term ‘good’ is admittedly vague; however, my point would be just as valid if the consequent were changed to \textit{I would be as good at basketball as Mark is} or \textit{I would make the first team}.
and the resultant differences in the laws that determine its shape, would be more significant than most other differences of fact in measuring overall similarity between worlds! I readily concede that I am not in any way ‘doing metaphysics’. I am solely motivated by epistemic and psychological concerns, and I do not think that maximal representations or omniscient agents have anything to do with normal counterfactual reasoning. I shall revisit this theme in chapter 4.

2.5.3 The Validity of Conditional Inference Rules

We have seen that conditional inference rules, including modus ponens, permit of counterexamples, but hold in many other cases. This would seem to show that these rules are not universally valid for natural language conditionals, i.e. a valid transitivity argument is not valid in virtue of being a transitivity argument, but in virtue of some other considerations. Whatever these ‘other considerations’ may be, they coincide with classical inference rules sufficiently frequently for those rules to remain explanatorily useful.

There is general acceptance that inference rules fail when implicit assumptions are shifted in the course of the argument. Walton (1989) alludes to the central role of vagueness in fallacious reasoning: “If vague terms are used in a consistent manner throughout the argument, there may be no logical difficulties or fallacies in the use of these terms” (p. 274). We have seen that vagueness is at the root of many problems involving counterfactuals, and it should be no surprise that it is also at the root of this one. In a possible worlds analysis, we can speak of shifting similarity metrics; in a metalinguistic analysis, we can speak of changing sets of relevant facts; in a situation semantics analysis, we can speak of changing background conditions.

When considering the shifting of background conditions, it is useful to introduce Lewis’s (1979) concept of conversational score. At any given point in a cooperative conversation, certain pragmatic parameters will be fixed to certain values; these values make up the score at that point in the conver-

\[\text{There are more than 10,000 students at TCD.}\]
\[\therefore \text{There are more than 2,000 students at TCD.}\]
sation\textsuperscript{27}. The set of presuppositions accepted at that point is an example of such a parameter; the reference of definite descriptions like \textit{your man over there with the hat} is another. The assumed boundaries of any vague terms that have been introduced are also part of the conversational score. Likewise the implicit information that licenses counterfactuals – the ordering of possible worlds, the set of relevant facts and beliefs, the background conditions, or whatever is preferred. When the speaker introduces a novel counterfactual scenario, e.g. \textit{If Bizet and Verdi had been compatriots, Bizet would have been Italian}, the hearer accommodates\textsuperscript{28} a mode of vagueness resolution that verifies that counterfactual. If the speaker then asserts a different counterfactual which requires a different mode of vagueness resolution, e.g. \textit{If Bizet and Verdi had been compatriots, Verdi would have been French}, then the conversational score must be altered.

Alterations of conversational score are not in themselves a problem. The score of any conversation will change continually, as presuppositions are added or cancelled, new, more salient entities enter the realm of discourse and take ownership of definite descriptions, and so on. What is crucial for the success of the conversation is that both parties follow these changes. One can felicitously assert that \textit{If Bizet and Verdi had been compatriots, Bizet would have been Italian}, and then that \textit{If Bizet and Verdi had been compatriots, Verdi would have been French}, but one cannot assert that \textit{If Bizet and Verdi had been compatriots, then Bizet would have been Italian and Verdi would have been French} on the basis of the acceptability of the preceding assertions. If counterfactual \textit{A} is true by some mode of vagueness resolution, and counterfactual \textit{B} is true by another, then it is generally fallacious to posit any relationship between the facts of \textit{A} and the facts of \textit{B} by either of those resolution modes. I would explain my dislike of the Conditional Excluded

\textsuperscript{27}There are obvious similarities between the idea of conversational score and that of common ground. In fact, depending on their formulation, it seems that either concept might well include, be part of, or be equivalent to the other.

\textsuperscript{28}Accommodation is a technical term for Lewis, referring to a process by which new information is incorporated into the conversational score. In general, “If at time \textit{t} something is said that requires component \textit{s\textsubscript{n}} of conversational score to have a value in the range \textit{r} if what is said is to be true, or otherwise acceptable; and if \textit{s\textsubscript{n}} does not have a value in the range \textit{r} just before \textit{t}; and if such-and-such further conditions hold; then at \textit{t} the score-component \textit{s\textsubscript{n}} takes some value in the range \textit{r}” (p 240 in Lewis 1983). In the case of counterfactual vagueness, one component of the relevant ‘such-and-such further conditions’ would presumably be the trade-off between revision of prior beliefs and the desire to satisfy the counterfactual.
Middle example in (16) in similar terms.

It should now be clear why the Hypothetical Syllogism fails. It does so when the implicit information that licenses one conditional premise clashes with the implicit information licensing the other. However, the information that is relevant in this case is not the same as that which was relevant in the previous case. The premises will very often be evaluated with regard to the same similarity criteria; what we must appeal to here is the strictness of vagueness resolution. In Lewis’ terms, strictness corresponds to the size of the sphere with respect to which a counterfactual holds. To alter the measure of strictness while discussing a counterfactual scenario is to alter the conversational score. If the degree of strictness is altered in the course of an argument, then one cannot be certain that the argument is valid; it may still be valid, but this will not be due to the argument form, merely to circumstance. Adams’ election example (11) demonstrates this. From a possible worlds perspective: in the closest worlds where Smith dies before the election, Brown will win, and in the closest worlds where Brown wins the election, Smith will retire to private life, but the closest worlds where Brown wins are closer than the closest worlds where Smith dies, and in those worlds Smith does not die. We can transfer the strictness condition to other theories: in a situation semantics, we can require that the background conditions licensing both conditional premise be equivalent, and also hold for the conclusion; this is known as the Conditional Xerox Principle: if $S_1 \Rightarrow S_2 \mid B$ and $S_2 \Rightarrow S_3 \mid B$, then $S_1 \Rightarrow S_3 \mid B$ (Seligman & Moss 1997: 303). In a metalinguistic framework, we require that the expanded antecedents of both premises are cotenable. Lewis (1973: 35) suggests a valid substitute for the Hypothetical Syllogism that essentially respects the strictness constraint:

$$
(24) \qquad \begin{array}{c}
\chi \Box \rightarrow \phi \\
\chi \land \phi \Box \rightarrow \psi \\
\therefore \chi \Box \rightarrow \psi
\end{array}
$$

In a similar vein, we can see that Antecedent Strengthening is only guaranteed to succeed when the sphere licensing the conditional premise contains at least one world where the strengthened antecedent holds. If the strengthened antecedent is less likely to hold than the antecedent of the conditional premise together with the negation of the strengthening proposition, as in (25), then the Antecedent Strengthening inference will be fallacious.

---

29. The diagrams in Lewis (1973: 34) are useful for grasping examples like these.

30. Though the conclusion may of course turn out true. For example, the addressee in
(25) If I had bought you a beer, you would be grateful. Therefore, if I had bought you a beer and then stolen your car, you would be grateful.

In metalinguistic terms, the strengthening proposition must be cotenable with the expanded antecedent of the conditional premise, which in the case of (25) presumably includes a proposition like *I do not commit any crime against you*. In situation semantics, the strengthening proposition must be compatible with both the situation corresponding to the antecedent of the conditional premise and the situation corresponding to the background conditions of the conditional premise. We can formulate a valid inference rule analogous to (24):

\[
\begin{align*}
\phi & \rightarrow \psi \\
\phi & \rightarrow \chi \\
\therefore \phi \land \chi & \rightarrow \psi
\end{align*}
\]

Contraposition fails for true-consequent conditionals like (10), and is generally unreliable when the consequent is significantly more probable than the antecedent. Failure of contraposition is shown in (27):

(27) If I had drawn a flush, I would not have drawn a straight flush. Therefore, if I had drawn a straight flush, I would not have drawn a flush. (Sanford 1989: 225)

One could argue that the premise contains a hidden scalar implicature, but no matter; were contraposition a valid inference rule for natural language counterfactuals, the argument would be valid in virtue of its form. It is difficult to specify when exactly contraposition succeeds without inviting circularity. The inference is valid when the antecedent of the premise is a sufficient condition for the consequent; however, part of the definition of a sufficient condition is that it licenses contraposition. Once again, the figure in Lewis (1973: 34) is useful in this regard.

Finally, we consider the so-called ‘conditional syllogisms’: modus ponens, modus tollens, affirming the consequent and denying the antecedent. Modus ponens has been described already; the other three are as follows:

(25) may be extremely rich, and might not care whether his car is stolen or not. The figure in Lewis (1973: 18) is illustrative in this regard.
On every interpretation of *if... then* that we have considered, modus ponens and modus tollens are valid inference rules; affirming the consequent and denying the antecedent are not. Proposed counterexamples to modus ponens, such as (12), violate conversational rules by blatantly shifting the conversational score; when this happens, validity cannot be guaranteed. Simple, two-premise instances of both modus ponens and modus tollens are always valid. Affirming the consequent and denying the antecedent are common fallacies; (28) illustrates a failure of affirming the consequent:

(28) If God had created the planets for use as snooker balls, the earth would not be flat. The earth is not flat. Therefore, God created the planets for use as snooker balls.

They are, however, valid when *if... then* statements are interpreted as biconditionals:

(29) If the light switch had been pressed, the light would be on. But the light is on. Therefore, the light switch was pressed.

It would seem to follow from my advocacy of a unified analysis of conditionals that inference rules should fail under the same circumstances for factual and counterfactual conditionals. There should be no valid argument involving indicative conditional premises that becomes invalid when these premises are replaced by the corresponding counterfactuals, and vice versa. Sanford (1989) suggests that (30a) seems valid, whereas (30b) seems invalid:

(30) a. If I am president of General Motors, I am very wealthy. If I am very wealthy, I drive a Jaguar. Therefore, if I am president of General Motors, I drive a Jaguar. (p. 225)

b. If I were president of General Motors, I would be very wealthy. If I were very wealthy, I would drive a Jaguar. Therefore, if I were president of General Motors, I would drive a Jaguar. (p. 226)
Sanford dodges the extreme oddity of (30a) by suggesting that the speaker has selective amnesia, which causes him to forget his identity, status and possessions. I admit that (30a) seems marginally more acceptable than (30b). In order to explain this, I must suggest that the presupposed falsity of the counterfactual antecedents has the effect of granting the hearer a greater degree of flexibility in manipulating similarity (sets of relevant facts, background conditions). Indeed, I believe that this is the case: given a conditional, the truth of whose antecedent is unknown, it seems reasonable to be more conservative in one’s background assumptions, as the antecedent may very well turn out to be true, and the conditional must then be evaluated in the actual world. Rott (1999), though arguing for a very different account of counterfactuals, makes a similar point about the comparative flexibility of factual and counterfactual conditionals.

2.5.4 Conclusion

We have now considered a number of philosophical theories of counterfactuals. It is fair to say that the possible worlds approach is the most widely known and accepted of these theories, but all share very many commonalities and congruences. No conclusive winner has emerged; indeed, as Lewis (1983: x) notes, it is extremely rare that one philosophical theory defeats another by sheer force of argument, and sometimes the only justification for championing a theory is its coincidence with our opinions. There is general agreement that the semantics of counterfactuals are straightforward, and that most problematic issues arise as a result of ‘pragmatic vagueness’. Attempts to tackle the pragmatic question of how to identify the appropriate information that licenses a given counterfactual have thus far made little progress. In even the best run of natural conversations, the speaker must find appropriate information on which to base his counterfactual assertions, and the hearer must decode that information, even though it is not explicitly conveyed. Humans thus cope surprisingly well with counterfactuals, yet philosophical analyses have proven far less successful. It may ultimately be the case, as Lewis suggests, that no sufficiently general account of the phenomenon can be found. This is, of course, no reason not to try.
Chapter 3

Psychological Perspectives
3.1 The Behavioural Psychology of Counterfactuals

3.1.1 How to Mutate Reality

The importance of counterfactual thought was first brought to the attention of psychologists by Kahneman & Tversky (1982). As part of their programme of investigation into the heuristics that drive people’s thoughts about the world, they identified a ‘simulation heuristic’. They proposed that activities such as prediction, probability assessments, counterfactual assessments and causal assessments are made by mentally manipulating aspects of reality, and using these as parameters to run ‘simulations’ of alternative events. Kahneman & Tversky also initiated the practice of studying the regularities in subjects’ mutations of reality.

Byrne (1996) identifies two functional properties of counterfactual scenarios: they are (i) recoverable and (ii) goal-driven. These properties distinguish it from unguided hypothetical speculation, and are integral to its potency. Recoverability dictates that people mutate aspects of reality in a way that ensures that the factual scenario is accessible from the counterfactual alternative being constructed. In general, the more significant the mutations made in creating a counterfactual scenario, the less similar and recoverable it is. As the purpose of counterfactual thought is usually a comparison with the factual world, the alterations to the status quo made by people thinking counterfactually tend to be as minimal as possible. Counterfactual scenarios are often (but not always) goal-driven, in that they are created for the purpose of producing a specific outcome (other than the factual one), or of establishing whether a different set of premises would lead to an outcome similar or different to the factual one.

A central question in the modeling of counterfactual thought asks which aspects of reality are the most mutable; this is a psychological version of the question of how individuals establish similarity across possible states of affairs. It is often observed (e.g. Byrne 2002) that people mutate reality in highly regular ways. Although we generally find it easy to think counterfactually, we seem predisposed to generating certain kinds of counterfactual situations and ignoring others. In the twenty years since Kahneman & Tversky’s pioneering work, a multitude experiments have been carried out by cognitive and and social psychologists trying to understand the ways in which people
create counterfactual scenarios. A number of ‘mutability effects’ have been identified by researchers; some have been shown to hold reliably in different circumstances, others are not even unanimously agreed to exist.

One robust finding is that small mutations of reality are preferred to large, complex mutations. Kahneman & Tversky (1982) told the story of two men, Mr. Crane and Mr. Tees, who both arrive at the airport 30 minutes after their respective flights were to take off. However, Mr. Tees’ flight had been delayed by 25 minutes, whereas Mr Crane’s had departed on time. 96% of subjects thought that Mr. Tees would be more upset than Mr. Crane, on the basis that it was easier to imagine how he could have been on time for his flight. It is a familiar experience that narrow defeats cause us more grief than failures which somehow seem inevitable.

A related observation is that exceptional events are mutated more frequently than ‘normal’ events, and that exceptional events are deleted more frequently than they are introduced. A second experiment by Kahneman & Tversky (1982) saw subjects presented with two different versions of a story about Mr. Jones, who is killed in a traffic accident while on his way home from work. In one version of the story, Mr. Jones takes a route other than his normal one; in the other version, he leaves work earlier than usual. The subjects were invited to suggest how the tragic accident could have been avoided. The results clearly demonstrated that increasing the abnormality of an aspect of the story made that aspect far more available for undoing; aspects that remained the same in both versions of the story tended to be just as mutable for both stories. The table below gives the percentage of replies for each story that suggested mutating one of the variables under investigation.

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Exceptional Time</th>
<th>Exceptional Route</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route</td>
<td>8</td>
<td>33</td>
</tr>
<tr>
<td>Time</td>
<td>16</td>
<td>2</td>
</tr>
</tbody>
</table>

Almost no responses suggested that an exceptional event be introduced, e.g. that he leave work earlier or later than normal in the Exceptional Route story. The authors conclude: “The psychological distance from an exception to the norm it violates is smaller than the distance from the norm to the same exception.” A similar experiment by Wells et al. (1987) confirms these findings.

Kahneman & Tversky (1982) also observed that actions seem to be more mutable than inactions; accordingly, people regret their actions more than
their failures to act after an undesirable outcome, and people usually feel a greater need to apologise for their actions than for their inactions. Further investigations by a number of authors have shown that this action effect is far from monolithic. A comparison by Gleicher et al. (1990) indicated that the action effect is more frequently observable in cases of regret (about bad outcomes) than in cases of satisfaction (about good outcomes). Furthermore, the action effect does not seem to obtain when people take a long-term perspective on events; when thinking about what they regret in life, people tend to wish that they had acted when they did not more than they wish they had not acted when they did. Gilovich & Medvec (1994) presented subjects with the following story:

Dave and Jim do not know each other but both are enrolled at the same elite East Coast university. Both are only moderately satisfied where they are and both are considering transferring to another prestigious school. Each agonizes over the decision, going back and forth between thinking he is going to stay and thinking he will leave. They ultimately make different decisions: Dave opts to stay where he is, and Jim decides to transfer. Suppose their decisions turn out badly for both of them: Dave still doesn’t like it where he is and wishes he had transferred, and Jim doesn’t like his new environment and wishes he had stayed.

In response to the question “Who do you think would regret his decision more upon learning that it was a mistake?”, 76% of subjects thought that Jim, who had transferred, would be most regretful. But in response to the question “Who do you think would regret his decision more in the long run?”, 63% of subjects named Dave, who had not acted. Byrne (1997) found that this reversal does not arise when subjects are asked to consider good outcomes; whereas people often relativise misadventures and get on with their lives, they remain happy about actions that turned out satisfactorily.

Wells & Gavanski (1989) investigated the influence of the presentation of experimental scenarios on counterfactual thought. In one experiment, they told a story about a woman, Karen, who was brought to a restaurant by her boss to celebrate her promotion. The boss was unaware that Karen had a severe allergy against wine. The boss ordered a dish containing wine; Karen had an allergic reaction and died. Two versions of the story were presented: in the first, the boss originally considered a wine-free dish; in the second, the boss originally considered a dish that also contained wine. Subjects were
asked to list four ways Karen’s death could have been avoided, to list the four most important causes of Karen’s death, and to rate the degree to which her boss’ decisions caused her death. Subjects given the first (initially wine-free) story were twice as likely to suggest a different choice of meal as a way of avoiding Karen’s death; they were almost twice as likely to list the boss’ choice of meal as an important cause of her death; and they rated the causal role of the boss as 50% higher. A second, similar experiment replicated these results. The authors conclude that the mutability of a given aspect of reality is greater when it is clear that changing it would have desired consequences. Such aspects correspond to Kahneman & Tversky’s (1982) joints of reality, events that are low in redundancy and high in causal significance. Furthermore, Wells & Gavanski’s experiments show that explicit descriptions of alternative scenarios and their consequences can make those scenarios more available.

A number of recent studies have sought to throw light on the Temporal Effect, by which later events in a temporal sequence are more likely to be undone than are earlier events. Miller & Gunasegaram (1990) used the following scenario to establish the existence of this effect:

Imagine two individuals (Jones and Cooper) who are offered the following very attractive proposition. Each individual is asked to toss a coin. If the two coins come up the same (both heads or both tails), each individual wins $1,000. However, if the two coins do not come up the same, neither individual wins anything. Jones goes first and tosses a head; Cooper goes next and tosses a tail. Thus, the outcome is that neither individual wins anything. From a logical point of view, neither Jones nor Cooper is more responsible than the other for the failure to win. However, Miller & Gunasegaram found that 86% of subjects predicted Cooper would experience more guilt; 92% predicted Jones would blame Cooper more than Cooper would Jones; and 89% chose to modify the scenario by mutating the second event. Segura et al. (2002) have shown that the Temporal Effect obtains for sequences of four events as well as sequences of two events. Byrne et al. (2000) demonstrated that it is the actual order of events that is crucial, not the order in which the sequence is presented (e.g. Jones tossed heads before Cooper tossed tails vs. Cooper tossed tails after Jones tossed heads). Sherman & McConnell (1996) invoke the Temporal Effect to account for the common experience that the final games in a basketball (equivalently football, hurling, cricket) season are
considered more important and decisive than earlier games, despite the fact that each game is equally important.

The standard explanation of the Temporal Effect states that the first event is instrumental in somehow initialising a simulation or mental model of the scenario, and that subsequent events are interpreted relative to that event. As a result, mutating the first event would compromise the integrity of the simulation/mental model. However, the immutability of the first event can be cancelled in a number of ways. Walsh & Byrne (2001) showed that changing the description of the experimental scenario could reverse the temporal order effect. The authors used a story similar to that of Miller & Gunasegaram (1990), describing a game where the two participants (John and Michael) must each choose a playing card from a deck. If the participants choose cards of different colours (i.e. one red, one black), then each wins 1,000; otherwise they win nothing. In their experiment, subjects were given different formulations of the winning conditions, or ‘counterfactual context’. The ‘red disjunction’ condition stated that “If one or the other but not both pick a card from a red suit, each individual wins 1,000”; the ‘black disjunction’ condition stated that “If one or the other but not both pick a card from a black suit, each individual wins 1,000”. In both versions, the facts remained the same: John goes first and selects a black card, then Michael selects another black card. When asked how the participants could have won, subjects who had been given the ‘black disjunction’ description acted in accordance with the temporal order effect, and counterfactually undid Michael’s selection, but subjects who had been given the ‘red disjunction’ description preferred to mutate the first event, i.e. John’s selection.

The tendency to undo more recent events in a temporal sequence can also be reversed when there is a strong causal link between the events. Given a causal chain $A \rightarrow B \rightarrow C \rightarrow D$, Wells et al. (1987) suggest that it is easiest to undo event $A$ as it is not seen as being causally determined by another event; the strength of this ‘causal effect’ depends on the strength of the causal relation. The second event in the sequence he pulled the trigger; the gun fired is intuitively immutable, whereas the first is not affected by the same causal constraints and can readily be undone. In an experiment, the authors related the story of a man driving across town to arrive at a sale before it ends; on his route, he encounters four hindrances (a flat tyre, a speeding ticket, a traffic jam and a group of senior citizens crossing the road), and arrives too late. It is implied that each event led to the next (because of the flat tyre, he had to exceed the speed limit, and so on). Irrespective of the ordering of the
four events, it was found that subjects preferred to undo the first event than subsequent events. Segura et al. (2002) confirm the existence of the causal effect, and demonstrate that it is independent of the number of events in the sequence.

3.1.2 Reasoning with Counterfactuals

‘As was observed with regard to their philosophical study, interest in reasoning with counterfactuals springs from a longer established investigation of indicative conditionals\(^1\). The central experimental paradigm in conditional reasoning, and possibly the most-studied experiment in psychology, is the Selection Task (Wason 1966). In this task, the subject is shown four cards face-down with either a number or a letter on the back. A typical set-up would look like this:

\[
\begin{array}{ccc}
A & 2 & P & 3 \\
\end{array}
\]

The subject is then given a rule, e.g. *If a card has a vowel on one side then it has an even number on the other side*, and asked which cards he would turn over in order to determine whether the rule is true or false. From a logical point of view, the task is trivial; the A and 3 cards should be selected. Yet in general, naïve subjects\(^2\) fail at the Selection Task. Wason found that the most frequent selections were cards displaying vowels and even numbers; almost no subjects selected cards displaying odd numbers. This core result has been replicated again and again; subjects generally (over 90% of the time, according to Griggs & Cox 1982) do not select the card displaying the negation of the rule’s consequent. Many subjects continue to fail at the task even after it has been explained to them.

Subsequent studies have shown that the nature of the rule to be tested has a dramatic impact on results in the Selection Task. In contrast to the dismal results obtained when abstract material is involved, rules with ‘realistic’ or ‘thematic’ content have elicited far better performance. Griggs & Cox (1982) found that subjects were more successful at testing rules such as *If a person is drinking beer then the person must be over 19 years of age* (corresponding to Florida state law at the time). A large body of evidence has confirmed that deontic rules or ‘social contracts’, involving concepts of permission and

\(^1\)Though only by the chronology of cognitive psychology could it be called ‘ancient’.

\(^2\)By naïve subjects, I mean subjects with no formal training in logic.
obligation, facilitate superior performance in the Selection Task (e.g. Cosmides 1989, Holyoak & Cheng 1995). Gigerenzer & Hug (1992) found that differences in ‘social perspective’ influence performance. Subjects were given the rule *If the hiker stays overnight, he must bring fuel*, and assigned either a ‘policing perspective’, whereby they were asked to test whether the rule was being observed, or an ‘epistemic perspective’, whereby they were asked to test whether the rule was true or false. Subjects adopting the policing perspective were substantially better at the task than those adopting the epistemic perspective. One can argue that the two perspectives correspond to two quite different versions of the Selection Task. The epistemic version is Wason’s version, and illustrates the verificationist bias in hypothesis testing; the data are taken as fixed, and the rule is tested accordingly. The policing version, on the other hand, takes the rule as fixed, and individual instances are checked against it.

Stenning & van Lambalgen (2002) argue on a similar basis that experimenters underestimate the diversity of effects in the Selection Task. In a qualitative study, they interviewed subjects performing the task, and found that subjects’ thought processes and comments differed significantly: some assumed the rule true, and selected cards that would verify it, some attempted to falsify the rule but chose inappropriate cards, some decided just to select the cards mentioned in the rule, some confused the meaning of *even* and *odd*, some professed utter confusion and could not justify their selections. Their conclusion is not that a formal analysis of the task is doomed, rather that the problem should not be oversimplified.

Another important, and for our purposes more relevant, paradigm in conditional reasoning studies is what might be called the inferential paradigm. Subjects are given a set of premises and asked what, if anything, follows from them. One robust finding of such experiments is that modus ponens (MP) inferences tend to be extremely easy, whereas modus tollens (MT) inferences are extremely difficult for most subjects. The normatively fallacious denying the antecedent (DA) and affirming the consequent (AC) inferences are made with intermediate regularity. Johnson-Laird & Byrne (1991: 53) report an experiment in which 97% of subjects successfully drew modus ponens inferences from a conditional such as *If Tony is in Kerry then Noel is in Dublin*; only 38% successfully drew modus tollens inferences. When given the categorical premise *Noel is not in Dublin*, the majority of subjects conclude that ‘nothing follows’. However, the same experiment showed that modus tollens inferences were significantly easier (59% success rate) from biconditional
premises such as *If and only if Tony is in Kerry then Noel is in Dublin*.

Byrne & Tasso (1994) showed that modus tollens is much easier from a counterfactual premise than from a factual conditional premise. The table below gives the percentage of correct responses to conditional syllogism tasks involving factuals and counterfactuals:

<table>
<thead>
<tr>
<th></th>
<th>MP</th>
<th>MT</th>
<th>DA</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual</td>
<td>100</td>
<td>40</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>90</td>
<td>80</td>
<td>80</td>
<td>50</td>
</tr>
</tbody>
</table>

We can see that every kind of inference besides modus ponens was drawn more often from the counterfactual premises. Modus ponens inferences were slightly less common in the counterfactual cases. The core findings of this experiment have been replicated in numerous studies, e.g. Byrne & Tasso (1999), Thompson & Byrne (2002).

As in the Selection Task, the content of experimental premises can have a dramatic influence on the inferences that subjects draw from them. Thompson & Byrne (2002) demonstrate that indicative conditionals which are interpreted as biconditionals due to their content (e.g. *If the butter was heated, then it melted*) elicit the same inferences as surface biconditionals, and that indicative conditionals which are interpreted as counterfactuals elicit the same inferences as normal subjunctive counterfactuals. An experiment by Johnson-Laird & Byrne (2002) shows how world knowledge can suppress modus tollens inferences; they gave subjects sets of premises based on geographical ‘inclusions’ or ‘exclusions’:

**Inclusion:** If Bill is in Rio de Janeiro then he is in Brazil
Bill is not in Brazil
∴ Bill is not in Rio de Janeiro

**Exclusion:** If Bill is in Brazil then he is not in Rio de Janeiro
Bill is in Rio de Janeiro
∴ Bill is not in Brazil

---

3A meta-survey of results from 23 conditional reasoning experiments is given in O’Brien et al. (1998: 394); it shows that performance can vary wildly across tasks. For example, proportions of correct modus tollens responses range from 12% to 87%. Although it is not made clear, we must assume that the experiments also differed significantly in design and in type of content, as the pattern found by Byrne & Tasso (1994) has been confirmed by many subsequent studies.
92% of subjects answered the inclusion problems correctly; only 34% answered the exclusion problems correctly. When the locations mentioned in the premises were replaced with unfamiliar placenames, this effect was reduced (although the inclusion problems remained easier than the exclusion problems).

Byrne (1989) has shown that even normally reliable modus ponens inferences can be ‘suppressed’. Subjects given pairs of conditional premises such as (i) and a categorical premise such as (ii) tend not to draw the logically valid conclusion “Lisa has a fish supper”.

(i) If Lisa goes fishing, then she has a fish supper.
   If Lisa catches some fish, then she has a fish supper.

(ii) Lisa goes fishing.

It would appear that the second conditional premise interferes with the first by adding a further condition; it is not sufficient that Lisa goes fishing in order for her to eat a fish supper, she must also catch some fish. When told (ii), subjects do not think that all the relevant conditions have been fulfilled, and generally conclude that ‘nothing follows’. Byrne et al. (1999) have shown that all four conditional syllogisms can be suppressed through the provision of additional antecedents.

Byrne & Walsh (2002) studied how individuals revise deductive conclusions when they encounter contradictory evidence. Subjects were invited to make inferences from conditional-categorical premise sets; they were then given factual information that contradicted their inferences, and asked to list what they still thought true. Two types of belief revision problem were used, as shown below:

<table>
<thead>
<tr>
<th></th>
<th>Modus Ponens</th>
<th>Modus Tollens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conditional:</td>
<td>$A \rightarrow B$</td>
<td>$A \rightarrow B$</td>
</tr>
<tr>
<td>2. Categorical:</td>
<td>$A$</td>
<td>$\neg B$</td>
</tr>
<tr>
<td>3. Conclusion:</td>
<td>$B$</td>
<td>$\neg A$</td>
</tr>
<tr>
<td>4. Contradiction:</td>
<td>$\neg B$</td>
<td>$A$</td>
</tr>
</tbody>
</table>

The authors investigated the existence or otherwise of the *Inference Contradiction Effect*. This effect had been proposed by Elio (1997), who surmised that reasoners prefer to revise their belief in the conditional premise when a modus tollens conclusion is contradicted, but prefer to revise their belief in
the categorical when a modus ponens conclusion is contradicted, even though the counterexample in both cases is the same \((A \land \neg B)\). Byrne & Walsh’s results confirmed the existence of the Inference Contradiction Effect. They also found that while factual conditional premises are on the whole more frequently revised than the associated categoricals, counterfactual premises are less likely to be revised. Counterfactual premises were almost never (16%) revised in the face of modus ponens contradictions; subjects were still more likely to revise counterfactual premises (38%) than categorical premises (13%) in response to the corresponding modus tollens contradiction, though an additional 29% responded by ‘affirming the contradiction and/or categorical’\(^4\).

3.2 A Theory of Reasoning: Mental Logic

Until recently, it was tacitly assumed that the ‘laws of thought’ with which we reason are the same as the laws of logic. That is, humans have an innate logical competence that corresponds to a set of formal truth-preserving rules. To claim otherwise, it was felt, would be to claim that man is irrational. Yet this thesis has a number of obvious flaws. Humans frequently appear to reason illogically. Logical deduction is infallible, yet humans regularly make fallacious inferences\(^5\); and if processes of inference are purely formal, then we cannot explain content effects like those observed in the Selection Task. However, Henle (1962) argues that our innate logical competence is indeed absolute, and that reasoners’ errors can always be attributed to misinterpretations of the premises.

Even if we were to accept that the laws of thought constitute a sound and complete logical system, it would still be necessary to establish the nature of that system. Might we think in modal logic, or linear logic, or relevance logic? Most likely, it would be some system as yet unknown. To claim, as Inhelder & Piaget (1958: 305) did, that “reasoning is nothing more than the propositional calculus itself” is clearly absurd, as many intuitively simple

\(^4\)Byrne & Walsh (2002) dispute that this is equivalent to revising the conditional premise (though it has the same effect). They observe that the inference from \(\neg(A \rightarrow B)\) to \(A \land \neg B\) and vice versa is very difficult for reasoners.

\(^5\)Of course, this by no means from a recent realisation. Frege (1979: 4) argued that “The laws in accordance with which we actually draw inferences are not to be identified with laws of valid inference; otherwise, we could never draw a wrong inference.”
concepts are inexpressible in classical first-order logic\(^6\).

Nevertheless, man is obviously adept at rational thought. Johnson-Laird & Byrne (1991) point out that an inherently irrational species could never have invented formal logic. On the other hand, a wholly rational species would never have needed to invent it. One mainstream view holds that our innate logical competence is accurate, but that cognitive processing constraints lead us to reason suboptimally; as Johnson-Laird & Byrne (1993b: 205) put it, that human reasoners “are rational in principle, but err in practice”. In this section and the next, I will consider two rival theories of reasoning that share this fundamental assumption.

Theories of mental logic are psychologically sophisticated versions of the ‘laws of thought’ hypothesis. They assume a core logical competence, in the form of a system of natural deduction, and posit cognitive constraints on the permissible complexity of deductions. The most prominent theory of this kind has been developed by Martin Braine, David O’Brien and their colleagues (e.g. Braine & O’Brien 1998a), and I shall concentrate on their work in this section. Another significant theory is that of Rips (1994).

The system of mental logic comprises rules of inference and a procedure for applying them. A representative inference schema is as follows:

\[
\begin{align*}
  p_1 \text{ or } \ldots \text{ or } p_n \\
  \text{Not } p_i \\
  \therefore p_1 \text{ or } \ldots \text{ or } p_{i-1} \text{ or } p_{i+1} \text{ or } \ldots \text{ or } p_n
\end{align*}
\]

In other words, if the negation of one disjunct of a disjunction is a premise, the remainder of the disjunction can be taken as a conclusion. In keeping with the formal nature of inference schemas, this conclusion will be valid irrespective of the content of the disjuncts. The schemas do not represent notions of truth or falsity, but they are truth-preserving; any legal inference from true premises will itself be true. Braine & O’Brien (1998b: 80–81) list the basic schemas assumed for their “mental-propositional logic”; these are divided into classes according to the conditions under which they may legally be applied.

The deductive process or ‘reasoning program’ is a set of routines (Braine & O’Brien 1999b). The core routine, the Direct Reasoning Routine (DRR) is itself executed in a number of stages. When making novel inferences from

\(^6\)For example, Barwise & Cooper (1981) prove that the meanings of quantifiers such as many and few cannot be defined in a first-order language.
premises, only one of these stages, the *Inference Procedure*, is required. This procedure seeks to match the premises under consideration to the input premises of one of the ‘core schemas’ (such as the one given above). The conclusion of the inference thus licensed is added to the set of premises, and the procedure executes again. If no premise matches the conditions of a core schema, additional ‘feeder schemas’ may be invoked. Feeder schemas are recursive inference schemas that would lead to infinite loops if their use were not restricted, e.g. $p_1, p_2, \ldots p_n; \therefore p_1 \land p_2 \land \ldots \land p_3$. Accordingly, they may only be used if their output is of the form required by a core schema\textsuperscript{7}, or when reading out conclusions. The procedure halts when a satisfactory conclusion has been inferred (success), or when no more inferences can be made (failure). In the latter case, the reasoner should conclude that ‘nothing follows’.

Evaluation of statements proceeds in a similar way, but involves two more components of the DRR. The *Preliminary Procedure* is invoked when a conditional proposition is to be evaluated; the antecedent is added to the set of premises, and the truth of the consequent is evaluated. The *Evaluation Procedure* then applies, as it does when testing non-conditional propositions: if the proposition is in the premise set, then the procedure returns “true”; if the proposition is incompatible with the premise set (as inferred using one of two ‘Incompatibility Schemas’ that yield contradictions, e.g. $p, \neg p \therefore \bot$), then the procedure returns “false; otherwise, the Inference Procedure is run, and the proposition is evaluated with regard to the new premise set. Once the Inference Procedure has been exhausted, the Evaluation Procedure must return “inconclusive”.

The inference schemas and the Direct Reasoning Routine are held to make up the totality of our innate logical competence, as shared by all cognitively developed adults. The DRR runs unconsciously; we have no direct introspective access to our inference schemas or the workings of the DRR procedures. The DRR is thus not goal-driven; it has not (to my knowledge) been made clear whether or not it is determinate. It is, however, extremely rapid and totally reliable. However, many reasoning problems cannot be solved using the

\textsuperscript{7}More accurately, if “one or a combination of the Feeder schemas” yield an appropriate premise (Braine & O’Brien 1998b: 82). As no limit is stipulated on the possible combinations of feeder schemas, it is unclear how this concession can avoid leading to overwhelming problems of intractability. One solution might be to rule out further application of feeder schemas during the unconscious DRR, but permit it as a goal-driven ‘Indirect Reasoning Strategy’, like reductios.
CHAPTER 3. PSYCHOLOGICAL PERSPECTIVES

DRR. For example, it does not license modus tollens inferences. Further ‘Indirect Reasoning Strategies’ are therefore required. One such strategy is that of reductio ad absurdum; another is the ‘Supposition of Alternatives’ (SAS): “If the premise set contains a disjunction... and if some of the propositions of the disjunction do not occur as antecedents of conditionals in the premise set, then suppose each of these in turn and try to derive a conditional with it as antecedent” (Braine & O’Brien 1998b: 82). Strategies are not universal, and are more difficult than the application of direct reasoning schemas. Strategies are learned, presumably implicitly; Bonatti (1998: 440) reports that SAS is common among college students.

The theory of mental logic makes predictions about the relative difficulty of reasoning tasks. Problems that can be solved using the DRR alone will generally be easy, though lengthy proofs may still result in errors; problems requiring indirect strategies will be more difficult. Differences in strategic competence will explain intersubject variation.

Braine & O’Brien (1991) outline a “theory of If” within the mental logic framework. The word if is treated as a logical connective for natural language. The authors propose that the lexical entry for if be identified with the inference schemas associated with that connective. There are two such schemas, which are as follows (p. 202–203):

1. **Modus Ponens**: if \( p \) then \( q \), \( p \); \( \therefore q \).

2. **Schema for Conditional Proof**: To derive or evaluate if \( p \) then..., first suppose \( p \); for any proposition \( q \), that follows from the supposition of \( p \) taken together with other information assumed, one may assert if \( p \), then \( q \).

The concept that underlies Schema 2 is familiar from our considerations of the metalinguistic philosophers. Neither the Schema or the logical reasoning program have anything to say about the nature of the “other information” to be assumed.

However, these Schemas are not enough to generate a plausible account of conditional reasoning. Schema 2 makes no reference to cotenability or any other constraint on the logical properties of “other information”. To this end, Braine & O’Brien posit a number of **Constraints on the Schema for Conditional Proof**:

1. Nothing follows from a contradiction except that some assumption is wrong.
2. A supposition can be the antecedent of a conditional conclusion reached via Schema 2 only if it is consistent with prior assumptions (i.e. premise assumptions plus any previously made suppositions).

3. An assumption reiterated into a conditional argument cannot contradict the supposition that is to be the antecedent of the conditional.

(p. 206)

Constraint 1 should be self-explanatory. The second constraint is essentially one of cotenability. The third constraint has a technical function, preventing non-cotenable premises from interfering with suppositions.

It follows from the Schema for Conditional Proof that conditionals with true consequents are always true, so long as the consequent does not contradict the antecedent, and that conditionals with false antecedents are neither true nor false. Given any premise \( q \) and any cotenable supposition \( p \), \( q \) will still hold under the supposition, and if \( p \), then \( q \) can be derived. Given a premise \( p \), it is not permissible to suppose Not \( p \), unless one is reasoning counterfactually.

The contraposition example below (from Braine & O’Brien 1991: 207) illustrates mental logic in action. The argument utilises the two conditional schemas as well as a reductio strategy:

1. If \( A \) then \( B \) (premise)
2. Suppose not \( B \)
3. Suppose \( A \)
4. If \( A \) then \( B \) (1, reiteration)
5. \( B \) (3,4, modus ponens)
6. Not \( B \) (2, reiteration)
7. Incompatible
8. \( \therefore \) Not \( A \) (reductio)
9. \( \therefore \) If not \( B \) then not \( A \) (Conditional Proof)

Mental logic can explain reasoners’ performance in conditional syllogism tasks. The almost universal success in modus ponens tasks is a consequence of its incorporation in one of the basic conditional inference schemas. Modus tollens, on the other hand, requires the use of an indirect reductio strategy, and should be far more difficult. Mental logic does not license denying the antecedent or affirming the consequent inferences. Yet as we have seen,
these inferences are drawn more frequently than are valid modus tollens inferences. O’Brien et al. (1998: 413) suggest that reasoners are susceptible to ‘invited inferences’, which are essentially generalised conversational implicatures. When reasoners cannot derive any useful conclusion from a set of premises, they may attempt to arrive at conclusions in accordance with pragmatic principles. A premise if $p$ then $q$ may ‘invite’ an inference if not $p$ then not $q$ or if $q$ then $p$, which yield DA and AC inferences respectively. It is unclear what (besides a clash with experimental data) rules out the invited inference if not $q$ then not $p$, which would lead to successful modus tollens inferences.

Braine & O’Brien (1991) describes how their theory can account for Byrne’s (1989) examples of MP-suppression. Reasoners interpret the two conditional premises If Lisa goes fishing, then she has a fish supper and If Lisa catches some fish, then she has a fish supper as having the meaning If Lisa goes fishing and she catches some fish, then she has a fish supper. It follows that the categorical premise Lisa goes fishing will not license the modus ponens conclusion. Vague ‘pragmatic principles’, based on world knowledge, are held to facilitate this particular interpretation (p. 214). Johnson-Laird & Byrne (1993a: 372) endorse an identical analysis of the phenomenon from the standpoint of mental model theory.

As might be expected given our earlier observations, Braine & O’Brien (1991) adopt an analysis of counterfactuals that borrows much from Goodman and Chisholm’s metalinguistic theories. The conditional inference schemas and constraints on the Schema for Conditional Proof also apply for counterfactuals, with the additional consideration that suppositions may contradict prior premises. To evaluate a counterfactual, a reasoner does as for a factual conditional: assume the antecedent, and evaluate the consequent. In this case, the set of premises must presumably be revised to permit the counterfactual antecedent (Braine & O’Brien do not attend to this). Much that was said of the metalinguistic theory in the previous chapter applies also to the mental logic theory of conditionals. Braine & O’Brien (1991: 219) recognise this commonality, and claim that invalid arguments like (25) can be explained in similar terms:

(25) If I had bought you a beer, you would be grateful. Therefore, if I had bought you a beer and then stolen your car, you would be grateful.

After we suppose the expanded antecedent, i.e. that I bought you a beer and then stole your car, it is no longer permissible to reiterate the condi-
tional premise, as it is no longer true. How do we decide that the conditional premise is no longer true? The authors are noncommittal as to how the glaring circularity in this argument can be avoided; it seems to me that they can refer to the “other information” (i.e. background conditions) that licensed the conditional premise, which should be contradicted by the supposition of the strengthened antecedent.

Opponents of mental logic theory (e.g. Byrne & Tasso 1999) have claimed that mental logic cannot explain the differences between factual and counterfactual conditionals with regard to the inferences that they license. In particular, it is argued that it cannot account for the fact that modus tollens is easier from a counterfactual premise. Indeed, counterfactuals do not provide any extra inference schemas from which modus tollens conclusions could be derived. As has already been seen, however, nonlogical considerations are assumed to impinge frequently on the domain of logical reasoning, and it is reasonable refer to such considerations in this instance. Unless he has evidence to the contrary, a reasoner will interpret a counterfactual premise like If John were Greek, he would live in Athens as presupposing the falsity of both antecedent and consequent. The categorical premise John does not live in Athens will confirm this presupposition, and the reasoner will feel confident in asserting that John is not Greek. The conditional premise If John is Greek, he lives in Athens does not signal any such presupposition, and the modus tollens inference requires a derivational procedure. We can explain the slight drop in modus ponens inferences from counterfactuals in a similar fashion: the categorical premise John is Greek will contradict the presupposition John is not Greek, and the reasoner may be less certain as to what follows. This analysis would also seem to predict increased denying the antecedent inferences from counterfactuals, on the basis that the categorical premise John is not Greek confirms the presupposition that the antecedent and consequent is false; the results of Byrne & Tasso (1994) verify this prediction.

It is not immediately clear that a similar argument can explain Byrne & Walsh’s (2002) findings regarding the revision of counterfactual premises. The modus ponens case seems tractable. Even when a counterfactual such as If John were Greek, he would live in Athens yields a modus ponens inference, i.e. given John is Greek, John lives in Athens is inferred, the presupposition that the antecedent and consequent are false might not disappear. Rather, they might cast doubt on the truth of the categorical premise and the conclusion. Upon learning that John does not live in Athens, reasoners might
feel that the presuppositions are being endorsed, and willingly abandon the
categorical premise. With modus tollens contradictions, it is less clearcut.
The premises *If John were Greek, he would live in Athens* and *John does not live in Athens* yield the conclusion *John is not Greek*, and the further information *John is Greek* leads to a contradiction. In this situation, the contradictory premise cancels the presupposition that the antecedent and the consequent are false; the validity of the modus tollens conclusion may be reconsidered, and with it the connection of the conditional premise to the other premises. It may then be relatively easy to revise the conditional premise, which seems somehow less concrete than the categorical. This ties in with Byrne & Walsh’s (2002) finding that very few subjects revised the categorical after modus tollens contradictions from either counterfactual or factual premises. Unfortunately, I do not see how the analysis can be extended to resolve the remaining explicandum, the observation that modus ponens contradictions from factual conditional premises tend to lead to revisions of the categorical premises.

The approach outlined in the previous two paragraphs has not (to the best of my knowledge) been proposed by adherents of mental logic theory, but it seems in keeping with its spirit. Crucially, they refer to situations in which the inferential machinery is bypassed in order to reach a conclusion; I do not see how the logical mechanisms proposed by the mental logic theorists can handle these problems. Mental logic is similarly incapable of providing a direct explanation of content effects in the Selection Task. Again, it must posit extralogical influences based on world knowledge that can hinder or facilitate conclusions (Braine & O’Brien 1991). Nor can mental logic explain the temporal or causal order effects, nor why exceptional events are mutated more often than normal events, nor why indicative conditionals are sometimes interpreted as biconditionals and sometimes not. Yet these failings do not entail a disaster for the theory; mental logic claims to give an analysis of human logical reasoning, not of representation. The influence of the interpretation and representation of premises on reasoning cannot be denied (see section 3.4 for discussion), but it is not the task of mental logic theory to account for it. Indeed, the main theses of mental logic theory would appear

---

8Where the conditional premise is factual, the lack of presuppositions regarding the truth or falsity of the antecedent and consequent should make revision of the conditional easier; indeed Byrne & Walsh (2002) found that modus tollens contradictions from factual premises led to revision of the conditional premise more often than did modus tollens contradictions from counterfactual premises.
to be that some (but not all, or even most) aspects of deduction are actually independent of representation, and that these aspects are handled by a reasoning program whose core components are universal.

3.3 Another Theory of Reasoning: Mental Models

The theory of mental models was devised as an alternative to accounts of reasoning that rely on formal rules (Johnson-Laird 1983). It holds that reasoning does not require a dedicated logical process, but rather emerges from the way facts and beliefs are represented in mental models. When we consider aspects of the world or sets of premises, when we imagine possible states of affairs, and when we comprehend dialogue, we create corresponding mental models, that represent their salient features and relations between those features. Accordingly, models may be constructed using perceptual, linguistic or otherwise symbolic information. They can represent temporal and spatial relations, abstract relations, logical annotations, and more. Models come in various sizes, from ones that represent a particular situation to the ‘small-scale model of reality’ that all humans carry around in their heads.

The proposal of a unified form of representation is intuitively appealing, not least because the idea of a ‘model’ is a familiar one. However, the term ‘model’ is ambiguous, and there is a danger of conflating different senses. McCarty (2003) observes that a model can be “either a simplified representation of something for purposes of study (model of), or a design for realizing something new (model for)”. Yu (2002) draws a distinction between models as duplicates or copies, as in ‘model ship’, and models as constructs sharing just some structural and/or functional characteristics of the original object, as in a computer simulation, or the twin-helix model of DNA. Regarding the fundamental properties of models, Johnson-Laird (1983: 403) writes:

The essential characteristic of a model is its functional role. A model is a high-level representation in what is, from a functional point of view, an arbitrary symbolic notation. The interpretive system treats an element in this notation, $A'$, as corresponding to an entity, $A$, in the world… There is no direct physical correspondence, because the pattern of neural events underlying $A'$ is quite different from the aggregate of molecules comprising $A$. 
There may be a structural resemblance, however, because the structure of a spatial model may be related to the corresponding physical structure.

It is clear from this description that mental models are to be understood as *models of*, in McCarty's terminology, and correspondingly as Yu's partial reconstructions. Elsewhere (e.g. Johnson-Laird 1983: ch. 1) the correspondence between the ‘relation-structure’ of a mental model and the ‘relation-structure’ of its referent in the actual world is stressed. The level of detail that is represented in a model will depend on the model's function and on the individual's knowledge of the world.

Accounts of the mental models theory of reasoning are quick to emphasise the distance between it and ‘formal’ theories such as that of Braine and O’Brien. Mental models has been described as a ‘semantic’ theory (e.g. Johnson-Laird & Byrne 1991: 23, Johnson-Laird & Byrne 1993b), as it eschews explicit inferential machinery and claims that conclusions can be drawn directly from what is represented in a model. The only processes employed in reasoning are those that construct models, a way of describing them, and a further check on the validity of conclusions.

There are three stages in the deductive reasoning process (Johnson-Laird & Byrne 1991: ch. 2):

1. Construction of a model (or models) of the state(s) of affairs described by the premises.

2. Formulation of a parsimonious description of the models that have been constructed (i.e. a conclusion).

3. Search for alternative models of the premises in which their conclusion is false.

The third stage is the key to ensuring the validity of deductions, as it requires that a conclusion be consistent with all states of affairs described by the premises. Fallacious reasoning results when we assert a conclusion without exhaustively checking all possible counterexamples. Cognitive resources are limited, and exhaustivity may not even be possible for complex problems. Mental models theory therefore makes a number of assumptions regarding the way in which human limitations affect reasoning:
1. The Principle of Truth (e.g. Johnson-Laird 2001): Mental models only represent true facts and beliefs by default; they do not represent falsehoods, nor states of affairs that are contradictory. If required, however, the falsity of a proposition can be established.

2. Mental models do not contain variables (Johnson-Laird & Byrne 1991: 212), nor can they directly represent large or infinite sets (Johnson-Laird 1983: 398). When we consider a sentence like There are 10,000 students at TCD, it is impossible to imagine 10,000 individual students; instead, we may construct a model containing a suitably large number of students, and proceed from there. If needs be, we can attach the annotation ‘10,000’ to the model to keep track of the number of students. If we are required to reason arithmetically from this premise, for example to calculate the effects of a 20% increase in student numbers, then we can depart from model-based reasoning and utilise our learned symbolic manipulation skills. Similarly, if a story begins Once there was a big fat alien from the planet Infundibulum 5, we tend to imagine a particular representative alien, probably based on our preconceptions of what aliens might look like, and retain this alien throughout our representation of the story. We do not interpret the opening line as ∃x(alien(x)∧big(x)∧fat(x)∧from(x,Infundibulum 5)) and represent the entire story within the scope of existential quantification over x. In syllogistic reasoning, the representation of a premise or conclusion All men are vegetarians cannot include every man; rather, a number of...

---

The concept is comparable to the use of Skolem constants in the programming language Prolog (cf. Clocksin & Mellish 1994:ch. 10).

Mental models theory provides an interesting analysis of ‘donkey sentences’ (so called because classic examples involve farmers and donkeys). Johnson-Laird & Byrne (2002: 659) give the following interpretation of the sentence If there is a car in the garage, then Mr. Toad will drive it:

- A car in garage
- Toad drives it
- ¬A car in garage

In the case where the antecedent is false, nothing is asserted about Toad driving anything, so quantificational problems are avoided. The pronoun it simply has no referent when there is no car in the garage; hence the impermissibility of If there isn’t a car in the garage, then Mr. Toad will drive it. Johnson-Laird & Byrne ignore the well-known problems of coreference that arise from such ‘donkey-anaphora’, and do not explain how the referent of it is assigned (other than by saying that it is). However, Garnham (2001) presents a mental models theory of anaphora resolution influenced by work in Discourse Representation Theory (Kamp & Reyle 1993).
arbitrary men are introduced in such a way that every salient possibility is represented (Johnson-Laird 1983: ch. 5).

3. The difficulty of a reasoning task depends on the number of models that it requires. If a problem requires the manipulation of many models, reasoners will not have the cognitive resources to generate them all, and any conclusions they might draw may be fallacious; alternatively, they may not observe any conclusion where one actually exists.

An example may illustrate how the theory works. Consider the following set of premises:

(a) John is happy or Jane is happy.
(b) Jane is not happy.

Under an inclusive disjunction interpretation, the first premise John is happy or Jane is happy yields the set of models\(^{11}\):

\[
\begin{align*}
\text{happy(john)} & \quad \text{happy(jane)} \\
\text{happy(john)} & \quad \text{happy(jane)} 
\end{align*}
\]

These models correspond to the possible states of affairs compatible with the premise, as determined by the meaning of the connective or – the meaning of a connective is just the set of models it supports. The second premise Jane is not happy yields a single model:

\[
\neg \text{happy(jane)}
\]

No violation of the Principle of Truth occurs here, as the token \(\neg \text{happy(jane)}\) is held to represent a true proposition. To formulate a conclusion, the two premises must be integrated. The second premise is only compatible with one model of the first premise, and we arrive at the integrated model:

\[
\text{happy(john)} \quad \neg \text{happy(jane)}
\]

\(^{11}\)The details of the model theory of propositional reasoning have varied slightly over the years. I have taken the specification in Johnson-Laird & Savary (1999) to represent the current state of the art.
Gricean principles prevent the repetition of premise information in the conclusion, so the appropriate deduction is that \textit{John is happy}.

The mental models account of conditional reasoning has undergone a number of revisions (cf. Johnson-Laird 1983: ch. 3, Johnson-Laird & Byrne 1991: ch. 4); the most recent version is presented in Johnson-Laird & Byrne (2002), and I will assume this version in the following discussion. Like mental logic, mental models theory offers a unified account of conditionals. Johnson-Laird & Byrne posit two ‘basic’ types of conditionals, and their default interpretations in the absence of content effects:

(a) \textit{If A then C}

Factual possibilities:

\begin{itemize}
  \item a c
  \item \neg a c
  \item \neg a \neg c
\end{itemize}

(b) \textit{If A then possibly C}

Factual possibilities:

\begin{itemize}
  \item a c
  \item a \neg c
  \item \neg a c
  \item \neg a \neg c
\end{itemize}

(a) represents the standard ‘conditional’ interpretation, which is compatible with any possible state of affairs other than those where \textit{A} is true and \textit{C} is false. (b) represents the ‘tautological’ interpretation, which the authors claim cannot be falsified\footnote{This may constitute a flaw in the model theory, as the statement \textit{If A then possibly C} would seem to be contradicted by the information that \textit{A and necessarily not C}. For example, if someone claims that \textit{If Elvis is still alive, then 2 might equal 3} (I see no distinction between \textit{possibly and might}), his conversational partner, who is both a mathematical realist and a conspiracy theorist, might deny the claim, on the basis that \textit{Elvis is alive, and it is impossible that 2 equals 3}. But the tautological interpretation remains unfalsifiable, as it is compatible with any state of affairs. The mental models account of modal reasoning, as presented by Bell & Johnson-Laird (1998), does not offer a solution – necessary falsity is identified with falsity in every model, and the integration of the two assertions thus yields the model set:

\begin{itemize}
  \item \text{Factual possibilities:}
  \item \text{alive(elvis) 2} \neq 3
  \item \text{\neg alive(elvis) 2} \neq 3
\end{itemize}

According to Johnson-Laird & Byrne, this set of models is still compatible with the assertion \textit{If Elvis is still alive, then 2 might equal 3}, although this contradicts any standard notion of possibility (even that of Bell & Johnson-Laird 1998). One would expect that if there is a model in which \text{alive(elvis)} is true, there must be at least one model where 2 = 3.
Johnson-Laird & Byrne’s (2002) proposal describes a number of fundamental principles that govern the representation of conditionals. The first is the Principle of Core Meanings: “The antecedent of a basic conditional describes a possibility, at least in part, and the consequent can occur in this possibility. Each of the two sorts of basic conditional accordingly has a core meaning referring to a set of factual or deontic possibilities” (p. 650). These core meanings are the sets of models given in the previous paragraph. It follows from this first principle that no necessary connection is assumed to hold between antecedent and consequent, although such connections can play a role in the interpretation of contentful (non-abstract) conditionals. Johnson-Laird & Byrne defend this consequence on the same grounds as Stalnaker (1968), viz. that conditionals may be true solely in virtue of their having true consequents, irrespective of any connection (p. 651). Furthermore, the Principle of Core Meanings does not invalidate the paradoxes of implication; the authors argue that reasoners’ aversion to these inferences is due to pragmatic principles that prevent the ‘throwing away’ of semantic information—it is more informative to say that Fred is not a dog than to say that If Fred is a dog, then Hannibal is a cat.

As previously observed, mental models theory offers a unified analysis of conditionals. Counterfactuals are treated similarly to other conditionals, as dictated by the Principle of Subjunctive Meanings: “A subjunctive conditional refers to the same set of possibilities as the corresponding indicative conditional, but the set consists either of factual possibilities or of a fact in which the antecedent and consequent did not occur and counterfactual possibilities in which they did occur” (p. 652–653). This principle heeds the fact that not all subjunctive conditionals are counterfactuals; a non-counterfactual subjunctive conditional yields the same interpretation as the corresponding indicative conditional. The (counterfactual) sentence If Fred were a dog, then Hannibal would be a canary thus yields the following set of models:

\[
\text{Fact:} \quad \neg \text{dog}(\text{fred}) \quad \neg \text{canary}(\text{hannibal})
\]
\[
\text{Counterfactual possibilities:} \quad \text{dog}(\text{fred}) \quad \text{canary}(\text{hannibal})
\]
\[
\neg \text{dog}(\text{fred}) \quad \text{canary}(\text{hannibal})
\]

is true. This situation is an unfortunate result of their wish to provide a maximally ‘basic’ core meaning for conditionals; it could be avoided by attributing some special status to the model where the antecedent and the proposition deemed possible by the antecedent are both true, as it is surely this model that licenses the conditional.
Similarly, the set of models for counterfactual *If A then possibly C* consists of one factual model with false antecedent and consequent, and three models representing counterfactual possibilities.

The meanings of conditionals are therefore unproblematic. However, reasoners do not always represent the meaning of a statement in its totality; working memory limitations lead them to represent only part of its meaning. They do not make all relevant models explicit. For example, reasoners frequently fail to consider the possibilities in which the antecedent of an indicative conditional is false. Rather, they just construct an initial set of models for a conditional premise, which may be fleshed out at a later point, should the need arise\textsuperscript{13}. The conditional *If Fred is a cat, then Hannibal is a cat* yields the initial models:

\[
\text{Factual possibilities: } \text{cat}(fred) \quad \text{cat}(hannibal)
\]

Our counterfactual example *If Fred were a dog, then Hannibal would be a canary* initially yields two explicit models:

\[
\begin{align*}
\text{Fact: } & \quad \neg \text{dog}(fred) \quad \neg \text{canary}(hannibal) \\
\text{Counterfactual possibilities: } & \quad \text{dog}(fred) \quad \text{canary}(hannibal)
\end{align*}
\]

The ‘\ldots’ in the above sets of models represent *implicit models*, whose existence is indicated, but not whose actual content is not. If needs be, the implicit models can be retrieved and explicitly represented (fleshed out). For a factual conditional, the implicit models represent the possibilities where the antecedent is false; for a counterfactual, there is only one implicit model, representing the possibility where the antecedent is false and the consequent is true.

This machinery is sufficient to account for subjects’ behaviour in basic conditional reasoning experiments. As we have seen, the initial and fleshed-out sets of models for a conditional premise *If there is a square, there is...*

\textsuperscript{13}Johnson-Laird & Byrne (1991) distinguish between *initial* and *fully explicit* or *fully fleshed-out sets of models*. In more recent writings the term *initial set of models* seems to have been replaced by the term *mental models*, e.g. “modus ponens can be drawn from the mental models of a conditional interpretation, whereas modus tollens can be drawn only from fully explicit models” (Johnson-Laird & Byrne 2002: 666). I find the adoption of this terminology needlessly confusing (are fleshed-out models not also mental models?), and have chosen to retain the earlier term *initial set of models*. 
a triangle are as shown in 1(a) and 1(b) respectively, and the initial and fleshed-out sets for a counterfactual premise If there were a square, there would be a triangle are as shown in 2(a) and 2(b) respectively.

1(a) Factual possibilities:  □ △

2(a) Fact: ¬ □ ¬△

Counterfactual possibilities:

1(b) Factual possibilities:  □ △

2(b) Fact: ¬ □ ¬△

Counterfactual possibilities:

¬ □ ¬△

The interpretation of the categorical premise There is a square is compatible with the single initially explicit model of the factual conditional premise. The categorical premise is therefore compatible with the set of initial models, and no further fleshing-out is required to produce the modus ponens conclusion There is a triangle. In contrast, the premise There is not a square is not compatible with any model in the initial interpretation of the factual conditional premise; to facilitate a modus tollens inference, the implicit model in 1(a) above must be made fully explicit, as in 1(b). Reasoners often fail to consider all possible models, however, and may simply conclude that nothing follows from the premises. Modus tollens inferences are thus more difficult than modus ponens inferences. Reasoners fallaciously affirm the consequent when they not only fail to make explicit the implicit model, but also fail to consider the existence of the explicit model (which indicates that some possibilities have yet to be considered). They deny the antecedent when they flesh out the implicit models only partly, i.e. they neglect the possibility in which the antecedent is false and the consequent true.

This analysis (initially presented in Johnson-Laird & Byrne 1991) would seem to predict that denial of the antecedent should be harder, and therefore less common, than affirming the consequent, as it requires a greater number of models. Yet O’Brien et al. (1998: 393) claim there is no conclusive evidence in the literature that this is actually the case; the results of Byrne & Tasso (1994), for example, show no difference in the frequency of DA and AC. However, O’Brien et al.’s argument is undermined by their own meta-analysis
of conditional reasoning experiments (p. 394) – of the 23 studies cited, all but one show a greater frequency for AC than for DA, and in many cases the margin is significant. Byrne & Tasso (1999) have also found that DA inferences are drawn significantly more often than AC inferences.

The mental models analysis also accounts for subjects’ inferential behaviour in counterfactual reasoning experiments. As for the factual conditional premise, one initial model of *If there were a square, there would be a triangle* is compatible with the interpretation of *There is a square*, and the modus ponens inference can thus be derived. The slight decrease in the frequency of MP inferences from counterfactual premises in comparison to factual conditional premises (Byrne & Tasso 1994, Byrne & Tasso 1999) may be explained by the greater number of initial models that are yielded by counterfactuals.\(^\text{14}\). The heightened availability of DA, as found in the same studies, can be attributed to the fact that it too is facilitated by the initial model set. Byrne & Tasso’s (1994) findings also indicate that AC inferences are more common from counterfactuals than from factual conditionals (something which mental models theory would not be able to explain); however, subsequent experiments (in Byrne & Tasso 1999) have failed to corroborate this.

In representing real-world conditionals, which do not deal with abstract content and which usually relate to prior knowledge and beliefs, interpretations other than the ‘core meanings’ may be used. The influence of world knowledge on interpretation is encapsulated in the *Principle of Pragmatic Modulation*: “The context of a conditional depends on general knowledge in long-term memory and knowledge of the specific circumstances of its utterance. This context is normally represented in explicit models. These models can modulate the core interpretation of a conditional, taking precedence over contradictory models. They can add information to models, prevent the construction of otherwise feasible models, and aid the process of constructing fully explicit models” (Johnson-Laird & Byrne 2002: 659). Pragmatic modulation explains those author’s findings that modus tollens inferences can be facilitated when they correspond to world knowledge, e.g. in spatial inclusion examples like *If Bill is in Rio de Janeiro then he is in Brazil. Bill is not in Brazil. Therefore Bill is not in Rio de Janeiro*. The knowledge that Rio de Janeiro is in Brazil will somehow prime the representation of the categorical

\(^{14}\)A ‘belief bias’ effect of the sort I advocated in section 3.2 would also be consistent with the theory, but it has not been proposed in the literature.
premise as

            Bill    ¬In Rio    ¬In Brazil

That is, the possibility in which Bill is not in Brazil is one in which Bill is not in Rio (Johnson-Laird & Byrne 2002: 667).

It follows from the Principle of Pragmatic Modulation and the similar Principle of Semantic Modulation\(^{15}\) that the interpretation of a conditional is not merely dependent on its formal properties. Johnson-Laird & Byrne outline 10 distinct sets of models, corresponding to 10 different ways of interpreting a conditional. For example, *If you log on to the computer, then you may be able to receive e-mail* will usually be interpreted as an 'enabling' conditional (p. 661):

\[
\begin{array}{c|c|c}
\text{Factual possibilities:} & \text{log on} & \text{receive} \\
\text{ } & \text{log on} & \text{receive} \\
\text{ } & \neg \text{log on} & \text{receive}
\end{array}
\]

This interpretation is presumably facilitated by the use of modal *may* and our general knowledge about computers and e-mail. The conditional *If you’re interested in Vertigo, then it is on TV tonight* receives a ‘relevance’ interpretation, as the hearer knows that his interest has no bearing on the TV schedule (p. 662):

\[
\begin{array}{c|c|c}
\text{Factual possibilities:} & \text{interested} & \text{on TV} \\
\text{ } & \neg \text{interested} & \text{on TV}
\end{array}
\]

One assumes that a visitor from the 18th century, knowing nothing of television or the works of Alfred Hitchcock, might interpret the sentence differently, perhaps as a basic conditional.

Content effects can now be demystified by taking into account modulatory influences on interpretation. In a sense, Johnson-Laird & Byrne’s proposal is in agreement with Stenning & van Lambalgen’s (1999) claim that “meaning acts as a hidden variable in psychological theories of reasoning.” The suppression of modus ponens demonstrated by Byrne (1989) is accounted for by the Principle of Pragmatic Modulation; the two premises *If Lisa goes fishing,*

\(^{15}\)“The meanings of the antecedent and consequent, and coreferential links between these two clauses, can add information to models, prevent the construction of otherwise feasible models of the core meaning, and aid the process of constructing fully explicit models” (Johnson-Laird & Byrne 2002: 658).
then she has a fish supper and If Lisa catches some fish, then she has a fish supper are both interpreted as expressing necessary conditions for having a fish supper and integrated into the same model. Ultimately, this analysis is functionally equivalent to that of Braine & O’Brien (1991). A slightly different mental models account is given by Byrne et al. (1999), who suggest that suppression occurs due to the ready availability of a counterexample model where Lisa goes fishing but does not catch any fish. This may be due to the representation of the concept of necessary conditions.

We have seen that subjects perform better in the Selection Task when they must reason with realistic or otherwise thematic material. Familiar content like A letter is sealed only if it has a 5d. stamp on it (Johnson-Laird et al. 1972) can enhance performance in the task, as can deontic content, even for an abstract permission rule like If one is to take action A, then one must first satisfy precondition P (Holyoak & Cheng 1995). How exactly these modulatory effects are manifested is unclear; in the case of deontic content, reference must ultimately be made to knowledge such as Holyoak & Cheng’s (1995) Pragmatic Reasoning Schemas. This issue will be taken up more fully in section 3.4.

Byrne & Walsh (2002) provide a mental model account of the Inference Contradiction Effect. As has been shown, modus ponens inferences may be drawn from the initial set of models for a factual conditional premise, whereas modus tollens inferences require the initial model set to be fleshed out. When subjects are presented with the contradictory premise Not B after inferring B from If A then B and A, they flesh out the implicit models of the premises, and generate the model ¬A¬B. In view of this compatible model, they cast doubt on the categorical premise A, and may choose to revise it. Once a modus tollens inference has been made, however, the initial set of models has already been made fully explicit, and all models incompatible with the premises have been rejected. The contradictory premise A does not match any remaining model, so it is the conditional which must be revised. The authors write: “The essential revision principle may be that a contradiction can be incorporated into one of the possibilities compatible with the conditional, if these possibilities have not been thought about and eliminated

---

16Johnson-Laird (1983: 33) describes an interesting Selection Task experiment by E. Golding in which this sentence was presented to a group of English subjects old enough to remember the pre-decimal currency, and to a group too young to have ever encountered a 5d. stamp. The former group performed well at the task, whereas the latter group performed no better than with abstract content.
already.” It is suggested that counterfactuals are less often revised because they already presuppose that the antecedent and consequent is false. The details of the analysis for counterfactuals are not spelled out, but it seems to me that the analysis is wrong.

The familiar initial models for the premise *If Fred were a dog, then Hannibal would be a canary* are:

Fact: \( \neg \text{dog(fred)} \\neg \text{canary(hannibal)} \)

Counterfactual possibilities: dog(fred) canary(hannibal)

... 

The categorical premise *Fred is a dog* eliminates the factual model and facilitates the modus ponens inference *Hannibal is a canary*. The contradictory premise *Hannibal is not a canary* is not compatible with the sole remaining initial model (dog(fred) canary(hannibal)), and nor is it compatible with the implicit model, which represents the possibility \( \neg \text{dog(fred)} \\text{canary(hannibal)} \). Byrne & Walsh must then predict, as they did with regard to modus tollens revisions from factual conditionals, that the conditional premise be revised. A similar situation obtains for modus tollens contradictions from counterfactual premises. The categorical premise *Hannibal is not a canary* eliminates the single initially explicit counterfactual premise (dog(fred) canary(hannibal)). The contradictory premise *Fred is a dog* is compatible with neither the factual model nor the single implicit model. We should again predict that the conditional premise be revised. Mental model theory therefore seems to predict that counterfactual premises will be revised more often than factual conditional premises, in direct contradiction to Byrne & Walsh’s (2002) findings.

An analysis of the temporal order effect has been described by Byrne (1997), Walsh & Byrne (2001) and by Segura *et al.* (2002). We recall that this effect involves the heightened mutability of more recent events in an acausal temporal sequence. The temporality effect is generally held to result from the fact that the initial event in a temporal sequence somehow initialises or contextualises the mental model of the sequence; Byrne *et al.*’s (2000) results indicate that this contextualisation is not simply a result of the order in which information is processed. We also recall Walsh & Byrne’s scenario of the card game, where each participant must choose a card of a different

---

17In accordance with the procedure for integrating models given by Johnson-Laird & Savary (1999).
colour (red or black) in order for them both to win. It is suggested that a reasoner might construct the following initial set of models to represent a given game in which the players lose:

Factual: John black Michael black Lose
Counterfactual: John black Michael red Win

... 

The temporal effect arises because only some winning scenarios are made explicit. However, changing the initial description of the game’s winning conditions can change the way the initial models are represented.

A ‘black disjunction’ description (If one or the other but not both pick a card from a black suit, each individual wins 1,000) is interpreted as follows:

John black Win
Michael black Win

In the losing situation where John and Michael both draw black, the two drawing events will match the first model in the representation of the winning conditions, and the initial models given above will be generated to represent the game. A ‘mental footnote’ will prevent the integration of both factual models in the same winning condition model. This ‘black disjunction’ condition facilitates the temporal order effect.

A ‘red disjunction’ description (If one or the other but not both pick a card from a red suit, each individual wins 1,000), on the other hand, yields the interpretation:

John red Win
Michael red Win

These models do not contain an explicit match to the factual situation in which both players draw black cards; when a subject is asked to generate a ‘win’ counterfactual, he selects the first model of the winning conditions, which has not been integrated into the factual model, and chooses to undo the first event, thus reversing the temporal order effect. This explanation

\[\text{\textsuperscript{18}}\text{Mental footnotes} have been an integral but often overlooked part of mental models theory since their introduction by Johnson-Laird & Byrne (1991). In general, footnotes on models constrain the ways in which they can be fleshed out by specifying which tokens have been exhaustively represented. In the current example, footnotes will prevent the integration of Michael black in the first model of the winning conditions and the integration of John black in the second model.\]

\[\text{\textsuperscript{18}}\]
seems reasonable, but one might ask what (besides the apparent serial order) causes the first winning condition, rather than the second, to be activated. One might even expect that the second winning condition would be primed by the first event (John draws black), and that the counterfactual model \textit{John black Michael black} would be the most available, leading to a mutation of the most recent event. Walsh & Byrne’s (2001) suggestion is that when the first event does not match any of the initial ‘goal’ (winning condition) models, a model containing the negation of the first event (i.e. \textit{John red}) is automatically selected. However, it is not clear that there is any independent motivation for this suggestion.

Mental models have also been invoked in an explanation of the action effect (Byrne 1997). This effect is demonstrated by Gilovich & Medvec’s (1994) college scenario, in which two unhappy students contemplate moving to a different university, but only one does. Byrne suggests that the initial models of the scenario are as follows:

- \textbf{Dave} Factual: at Uni X unhappy
  Counterfactual: move to Uni Y happy
  move to Uni Y unhappy
- \textbf{Jim} Factual: at Uni X unhappy
  move to Uni Y unhappy
  Counterfactual: stay at Uni X happy

The corresponding fully explicit models are therefore:

- \textbf{Dave} Factual: at Uni X unhappy
  Counterfactual: move to Uni Y happy
  move to Uni Y unhappy
- \textbf{Jim} Factual: at Uni X unhappy
  move to Uni Y unhappy
  Counterfactual: stay at Uni X happy

It is easier to flesh out the set of models for Jim, as there is only one counterfactual model to be constructed (as opposed to two for Dave). Although the model in which Jim stays at University X and becomes happy is intuitively the least plausible of all, it is nevertheless the most available. An additional factor in its availability may be humans’ need to believe that unhappiness is not inevitable, and that there is always some way that things could be made better. Regarding the reversal of the action effect in long-term retrospection, Byrne writes: “We may feel compelled to leave no stone unturned and no counterfactual model unexamined in our attempt to see whether our lives could have been better” (p. 130).
Most experimental investigations of counterfactuals have either invited subjects to adopt a speaker’s perspective and generate novel counterfactuals, or invited them to adopt a mute hearer’s perspective and reason from premises that are not to be questioned. Byrne & Tasso’s (1999) study of verification and falsification of counterfactuals thus reflects a novel approach. The experimenters presented subjects with a description of a situation involving shapes drawn on a blackboard, and the rules *If there was a circle on the blackboard, there was a triangle and If there had been a circle on the blackboard, there would have been a triangle.* They then asked “what two shapes would best fit the description and what two shapes would definitely go against it” (p. 735). The results were as follows:

<table>
<thead>
<tr>
<th>Task</th>
<th>Shapes Generated</th>
<th>◦,△</th>
<th>not-◦,not-△</th>
<th>◦,not-△</th>
<th>not-◦,△</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifying:</td>
<td>Factual</td>
<td>78</td>
<td>17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Counterfactual</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Falsifying:</td>
<td>Factual</td>
<td>0</td>
<td>44</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Counterfactual</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

The results of the verifying tasks seem clear: the majority of subjects thought that the co-occurrence of a circle and a triangle would verify the factual conditional and that either the co-occurrence of a circle and a triangle or the absence of both circle and triangle would verify the counterfactual. From a mental models viewpoint, these correspond exactly to the predicted initial models. In the falsifying task, the responses for the factual conditional indicate a distinction between those who fully fleshed out their models and gave the normatively correct answer ◦, not-△ and those who failed to fully flesh out their models and answered not-◦, not-△. The falsifying responses for the counterfactual indicate that only 30% made their models fully explicit, and that the remainder were divided between those who chose to negate the factual model and those who chose to negate the single explicit counterfactual model. The mental models account thus accords neatly with the data.

What is striking is that verification in this case seems to have little to do with truth. Despite its claims to be a ‘semantic’ theory, the architects of mental models are wont to assert that mental models merely represent ‘possibilities’ (e.g. (Johnson-Laird 2001)), and that judgements of truth or falsity
involve a ‘meta-ability’ beyond the ability to match models with other models (Johnson-Laird 1983: ch. 10, Johnson-Laird & Byrne 2002: 652). Indeed, it is hard to imagine how a completely abstract counterfactual statement like *If there had been a circle, there would have been a triangle* can be judged true or false in the absence of any contextual information about circles or triangles. In possible-worlds terms, there is no basis on which circle-worlds can be compared for similarity. Johnson-Laird & Byrne write:

One cannot observe counterfactual states, and so the truth or falsity of a counterfactual conditional about a contingent matter may never be ascertained... Our concern is the meaning of conditionals, not their truth or falsity. (p. 652)

The results of Byrne & Tasso’s (1999) experiment constitute convincing evidence for the existence of a ‘core’ meaning of counterfactuals, even if it has little to do with any realistic aspect of their usage. Whether or not this core meaning is that provided by mental model theory has not been conclusively shown. The alternative ‘core meaning’ suggested by Braine & O’Brien (1991), comprising the conditional inference schemas, is nonetheless undermined by Byrne & Tasso’s results. The Schema for Conditional Proof does not license the verification of the counterfactual in the situation not-○,not-△, but it does license its verification in the situation ○,not-△.

Byrne & Tasso claim that their results rule out the possibility that subjects adopt a Lewisian approach to counterfactuals. In particular, they assert that a possible worlds-type analysis would predict that *If there had been a circle, there would have been a triangle* should be verified by a circle with a triangle, and falsified by a circle with no triangle. This assumes that a Lewisian subject would have interpreted the question as asking which possible situation would license the counterfactual’s assertion in the actual world. An alternative interpretation would be the question of which past actualised situation would have licensed the counterfactual’s assertion with respect to possible worlds.

---

19 Abstract counterfactuals such as this may appear problematic for my previous suggestion (in section 2.5) that the truth of the consequent can break ties in similarity between possible worlds. My proposed criterion would seem to render all such counterfactuals true, as no facts or beliefs will inhibit the acceptability of the consequent. However, I would claim that in a conversational situation, the hearer will accommodate any abstract counterfactual as true, though he may question the speaker’s reasons for asserting it. Verifying a counterfactual of this type, on the other hand, would seem nigh on impossible if one is not to resort to wild stipulation.
some non-actualised possible world(s). Under this interpretation, the possibility \( \neg \sigma, \triangle \) should have been selected at least as often as \( \neg \sigma, \neg \triangle \) to verify the counterfactual, and \( \sigma, \neg \triangle \) should have been selected at least as often as \( \neg \sigma, \neg \triangle \) to falsify it. The former prediction is refuted by the results of the experiment, although the latter prediction is endorsed. We conclude that this interpretation cannot have been the most common or primary one. Of course, one could also argue that it is nonsensical to ask questions about the verity of abstract counterfactuals, and that the scattered distribution of responses reflect a confusion as to what is actually required of them. The responses to the verification task may reflect a division among those whose interpretation of the question required an answer in which antecedent and consequent are true (i.e. a non-counterfactual interpretation) and those who chose to endorse the counterfactual presupposition that the antecedent and consequent had been false. In the falsification task, confused subjects may simply have negated their answer from the verification task, whereas other subjects may have made a Lewisian interpretation.

3.4 Discussion

3.4.1 Models or Rules?

Mental models is a unified theory of reasoning and representation; mental logic is merely a theory of reasoning. Therein lies the core difference between the two theories as regards spirit, if not detail. The following quote from Johnson-Laird (1983: 397) reflects the degree to which mental models is held to provide a cognitive ‘theory of everything’:

It is now plausible to suppose that mental models play a central and unifying role in representing objects, states of affairs, sequences of events, the way the world is, and the social and psychological actions of daily life.

Indeed, adherents of the theory have provided accounts of an impressive range of cognitive activities, including causal reasoning (Goldvarg & Johnson-Laird 2001), probabilistic reasoning (Johnson-Laird 1994, Johnson-Laird et al. 1999), modal reasoning (Bell & Johnson-Laird 1998), the comprehension of language and discourse (Garnham 1996, Garnham 2001), and, of course, deductive and counterfactual reasoning. Davis (1993) approvingly observes that mental models are more probable candidates for explaining reasoning abilities
in non-human animals; a number of authors (e.g. Johnson-Laird 2001, Yu 2002) have suggested that reasoning abilities developed from perceptual systems and not from systems underlying symbolic manipulation. Support for the theory has also come from researchers in the field of scientific creativity; Tweney (1993: 367) writes that “Virtually any scientific diary can be read as an articulation of a mental model”.

The theory of mental models is thus supported by a strong argument from parsimony. Mental logic theories have only been fully developed to account for propositional deductive reasoning tasks such as are rarely encountered outside of psychological experiments. The relative lack of generality afforded by mental logic is not necessarily an argument against that theory, as (a) the fact that the theory has not been extended into other areas of reasoning does not entail that it cannot, or will not be; (b) one cannot criticise a theory for not explaining phenomena that it has never claimed to explain; (c) the assumption that cognition is logical in nature is one shared with semanticists who posit a propositional ‘Logical Form’ as the underlying mode of sentence representation, and with proponents of the ‘language of thought’ hypothesis (e.g. Fodor 1975). Nonetheless, the relative generality afforded by mental model theory is significant evidence in its favour.

It is fairer to try to decide the issue on a battlefield that both sides have chosen, viz. propositional reasoning. Although a conclusive judgement lies outside the remit of the present study, it will be useful to consider briefly the state of the battle, especially where it relates to conditional and counterfactual reasoning. The typical shortcomings of mental logic, as identified by mental model theorists, include the following:

1. Mental logic cannot account for the results of basic inference tasks from conditional and counterfactual premises (Johnson-Laird & Byrne 1991, Byrne & Tasso 1999).

2. Mental logic cannot account for the results of ‘verifying situation’ experiments (Byrne & Tasso 1999, Johnson-Laird & Byrne 2002).


4. Mental logic cannot account for the temporal order or action effects in counterfactual generation experiments (Walsh & Byrne 2001, Byrne 1997), or for the inference contradiction effect (Byrne & Walsh 2002).
5. Mental logic cannot explain content effects in the Selection Task, or the effects of semantic and pragmatic modulation on inference and interpretation (Johnson-Laird & Byrne 2002).

We have already seen that objection 1 is no longer tenable once we take the natural step of rooting the system of mental logic in a broader system of representation. The differences between factual and counterfactual conditionals that have been identified by conditional syllogism tasks can be attributed to the counterfactual’s presupposition that its antecedent and consequent are both false.

Objection 2 is more interesting. The results of Byrne & Tasso’s (1999) ‘verifying situation’ experiment indicate that people verify statements by matching their possible interpretations to actual states of affairs, although one should be wary about the results of the corresponding falsification experiment due to the confusion it seemed to cause among subjects. Byrne & Tasso results strongly contradict Braine & O’Brien’s (1991) prediction that subjects will verify a counterfactual if and only if the supposition of the antecedent entails the truth of the consequent, just as is the case for factual conditionals. The results of a subsequent verification experiment by Johnson-Laird & Byrne (2002) pose additional problems for the mental logic theory. In this experiment, subjects were given three statements, and asked which states of affairs would make each statement true. As in Byrne & Tasso’s experiment, the statements had abstract content such as If there is an A on the board then there is a 2, and if there is a C on the board then there is a 3. The results (percentages of participants who generated each possibility) were as follows:

<table>
<thead>
<tr>
<th>Possibilities</th>
<th>If A then B, and C then D</th>
<th>If A then B, and C then D or C and D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>84</td>
<td>92</td>
</tr>
<tr>
<td>CD</td>
<td>80</td>
<td>96</td>
</tr>
<tr>
<td>ABCD</td>
<td>80</td>
<td>96</td>
</tr>
</tbody>
</table>

These results are a striking endorsement of mental model theory, which predicts that all three statement types yield the same set of initial models. It is not clear how mental logic can account for Johnson-Laird & Byrne’s findings. It should be noted, however, that a subsequent replication of the experiment (also reported in Johnson-Laird & Byrne 2002) provided less clearcut (but
similar) data. A falsification experiment was also performed with the same stimuli, but the authors report that subjects found the task of generating false possibilities extremely difficult, and tended merely to negate their verifying responses. This reflects the evidence from Byrne & Tasso (1999) that falsification tasks are too hard in many circumstances.

Objection 3 concerns ‘illusory inferences’, inferences that seem compellingly obvious, yet are invalid. Johnson-Laird & Savary (1999) provide the following example about a hand of cards:

If there is a king in the hand then there is an ace in the hand, or else if there is not a king in the hand then there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

Reasoners almost invariably infer that there is an ace in the hand. This inference is, however, not valid; in fact, there cannot be an ace in the hand. Intuitively, one accounts for the illusion by noting that one seems to forget the fact that one premise must be false. Johnson-Laird & Savary suggest that reasoners construct the following mental models of the first premise:

<table>
<thead>
<tr>
<th>king</th>
<th>ace</th>
</tr>
</thead>
<tbody>
<tr>
<td>¬king</td>
<td>ace</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Upon learning that there is a king in the hand, they automatically endorse the conclusion that there is also an ace.

Johnson-Laird & Byrne (2002) contend that mental logic cannot explain this ‘illusory inference’, as formal rule theories “rely solely on valid principles of inference, which cannot account for systematic invalidity” (p. 673). This may not actually be the case, however. Bonatti (1998) describes how invalid inferences can follow when reasoners fail to “remember whether a sentence comes from the application of a rule, is a temporary supposition, or comes from a rule within a temporary subderivation” (p. 444). Accordingly, one might forget that one of the disjuncts in the premise If there is a king in the hand then there is an ace in the hand, or else if there is not a king in the hand then there is an ace in the hand has merely been supposed for the sake of the argument, and treat inferences from it as true:
1. If K then A or else If not K then A (premise)
2. K (premise)
3. Suppose If K then A
4. K (2, reiteration)
5. A (3,4, modus ponens)
6. A (5, error)
7. Suppose If not K then A
8. Suppose not K
9. K (2, reiteration)
10. Incompatible
11. ∴ Nothing follows (constraint on Conditional Proof)
12. ∴ A

All that has gone wrong here is that the reasoner has neglected to close off his supposition. This is in keeping with the intuitive feeling that one has forgotten to consider the situations in which the disjuncts are false (as is the mental models account).

On objection 4: It is true that mental logic cannot account for the temporal order effect or for the action effect. These effects must emerge from a system of representation, not from a system of reasoning. Byrne’s (1997) argument that these effects arise when we do not consider all possible scenarios is very plausible. One might question what aspect of these effects necessitates an analysis in terms of mental models, unless we are to rather weakly identify the mental models analysis with the thesis that imaginative thinking involves the consideration of alternative states of affairs, and that we generally do not consider all possible states of affairs. Likewise, no distinctive property of mental models can explain the heightened mutability of exceptional events, or the causal order effect. These phenomena are presumably emergent properties of the mode of representation as well. One cannot argue that an account of such phenomena in terms of mental models validates the mental models theory of reasoning, which is the matter at hand. Mental logic cannot account for the results of Byrne & Walsh’s (2002) belief revision experiments, but we have seen that the mental models account is not without problems either.

I will briefly postpone consideration of objection 5, as much of the remainder of this section will be concerned with issues it raises. Before turning attention to these issues, it will be convenient to conclude, or at least put a temporary halt to, the technical dispute between mental models and mental
logic. We have covered many of mental model theory’s objections to mental logic; fairness dictates that we now discuss the arguments of the mental logicians.

One significant objection against mental models theory has found fault with the proposed mechanism of recording ‘exhaustivity’ in models. A mental footnote to a set of models can record that a given token has been exhaustively represented, and that that token cannot appear in any models that have not yet been fleshed out. Such footnotes play an important role in facilitating the fleshing-out process and the deduction of valid conclusions. Johnson-Laird & Byrne (1991) use a square-bracket notation to represent exhaustivity; the notation seems to have disappeared in much subsequent work, though the concept of exhaustive representation remains part of the theory. Using this notation, the initial models of a conditional If \( A \) then \( B \) are as follows:

\[
[A] \quad B
\]

\[
\ldots
\]

The footnote records that \( A \) has been exhaustively represented in the set of models, and will therefore not be true in any of the implicit models. O’Brien et al. (1998) claim that the concept of exhaustive representation is unworkable, in view of the fact that “Exhaustivity is a relation rather than a property, although models theory thus far has treated it as a property” (p. 405). To illustrate this argument, they offer the following example:

Premise 1: \( \text{if } p \text{ then } q \)

\[
[p] \quad q
\]

\[
\ldots
\]

Premise 2: \( \text{if } r \text{ then } q \)

\[
[r] \quad q
\]

\[
\ldots
\]

Combined models for 1 + 2:

\[
[p] \quad [r] \quad q
\]

\[
\ldots
\]

Premise 3: \( p \)

Combined models for 1 + 2 + 3:

\[
[p] \quad [r] \quad q
\]

A reasoner should therefore conclude \( r \) and \( q \), though this is clearly wrong. O’Brien et al. argue that exhaustivity cannot be treated as a binary attribute, but as a relation, so that the following hold:
CHAPTER 3. PSYCHOLOGICAL PERSPECTIVES

\[ p \] is exhausted with regard to \( q \)
\[ r \] is exhausted with regard to \( q \)
\( q \) is not exhausted with regard to \( p \) or \( r \)
\( p \) is not exhausted with regard to \( r \)
\( r \) is not exhausted with regard to \( p \)

This reconception would require a representation of relative exhaustivity:

\[ [p, q] [r, q] q \]

O’Brien et al. view this reconception as problematic, as it very rapidly leads to intractability, and makes reasoning tasks more difficult than they should be.

The problems with the above example seem to arise when Premises 1 and 2 are combined. We would like to avoid a set of models in which \( r \) only occurs when \( p \) occurs and vice versa, as this clearly does not follow from the premises. Rather, a model set such as follows would be more appropriate:

\[ [p] q \]
\[ [p] [r] q \]
\[ [r] q \]

In order to yield this model set, we must allow the model \([p] q\) to combine with the implicit model \(\ldots\) and produce the model \([p] q\), not the null model as the theory seems to suggest, and similarly with the model \([r] q\). Johnson-Laird & Savary (1999: 200) state that this integration can only occur when the explicit model contains no item in common with any model in the set of models in which the implicit model occurs. Yet the model set for \(if\ r\ then\ q\) has the item \(q\) in common with the explicit model, and the integration is blocked. This seems to be a problem for the model theory as currently stated.

It is clear that a theory of mental logic must be rooted in a theory of mental representation that explains content effects and visual reasoning. Its proponents even concede that this could be provided by mental models: “That people often use mental models in reasoning is entirely compatible with our proposal and we believe that a complete account of inference will need a subtheory of mental models” (Braine & O’Brien 1991: 231). What is less clear, and has been disputed, is whether mental model theory needs a theory of mental logic. Mental models and the processes underlying them certainly
implement a logic, containing elements of disjunction (sets of disjoint models), conjunction (the integration of models) and negation (explicit negation tags). A principle equivalent to the Law of Non-Contradiction is required to block the integration of models containing a token and its negation. Mental models are inherently logical; in this sense they differ from non-logical theories of reasoning, such as Oaksford & Chater’s (1998, 2001) theory of probabilistic reasoning. However, mental model theory holds that its logical principles are not implemented by means of formal rules of inference; rather, they are immanent in the mode of representation and the general principle of deduction which states that a conclusion is valid if no counterexample model exists.

Braine (1993) claims that mental models theory requires the use of variables and a rule \( \neg\neg x \equiv x \) to capture the notion of exhaustivity, i.e. that any token in an implicit model cannot be the same as a token which has been exhaustively represented in an explicit model. A related issue (implicit in Braine’s argument) is what ‘meaning’ mental models theory assigns to quantifiers like All, Some, Few; clearly they cannot ‘mean’ sets of models, and would indeed seem to require interpretations that refer to variables. These questions are addressed by Johnson-Laird & Byrne (1991: ch. 9), who posit an intermediate level of representation (essentially Logical Form) onto which syntactic structure maps, and which in turn maps onto conceptual structure (i.e. mental models). This level of representation does allow variables, and the mapping onto mental models introduces arbitrary representatives to obviate the need for variables. One could argue that this does not explain how exhaustivity is implemented on the model level; Johnson-Laird & Byrne would presumably claim that it, like negation, conjunction and disjunction, is part of the implicit a priori logical machinery that constructs models. One can imagine something like an ‘exhaustivity switch’ that, when flipped for a particular token, blocks further representations of that token. The function of the switch can be described using variables, but it does not explicitly use them.

So: can we now reach a verdict on the mental logic-mental models debate? We have considered four of five objections made by mental model theorists against mental logic (we will soon proceed to the fifth). Objection 1, relating to conditional syllogism tasks, was rejected. Objection 2, relating to verification tasks, was accepted, with the reservation that the corresponding falsification tasks seemed to lead to confusion rather than provide any conclusive evidence. Objection 3 held that mental logic cannot account for
‘illusory inferences’; I showed that mental logic can explain Johnson-Laird & Savary’s (1999) conditional illusory inference results, but similar illusions have been identified in other areas of reasoning and it is not obvious that a formal rule theory can account for all of them. Objection 4 against mental logic related to its inability to explain certain phenomena regarding the generation of counterfactuals, specifically the temporal order and action effects; I argued that mental logic is not a theory of representation, and should not be criticised for not doing what it has never claimed to do. I also pointed out that neither theory can account for the inference contradiction effect. We then considered some arguments against mental models theory. The concept of exhaustivity was seen to lead to problems and unacceptable predictions; a clarifying statement from the mental model theorists on the implementation of exhaustivity would be helpful in this regard. The argument that mental model theory requires a subtheory of mental logic was rejected; it undoubtedly implements logical concepts, just as do AND- and OR-gates, but no additional reasoning routine is necessary.

The debate is inconclusive, and will probably remain so. One may validly ask what it would take to falsify either theory. Johnson-Laird (2001: 441) states that mental model theory would be ‘overturned’ by any finding that a problem requiring many models was easier than a problem requiring few models. However, critics have argued that the procedures for quantifying of problem difficulty have not been made clear. Newstead (1993: 359), discussing the model account of syllogistic reasoning, writes of the “lingering suspicion that the number of models a syllogism permits is to some extent determined by how difficult it is found to be.” O’Brien et al. (1998) claim that different mental model studies have used different criteria to distinguish ‘easy’ inferences from ‘hard’. It is probably important to differentiate between the general ‘frameworks’ of mental models and mental logic, and individual ‘theories’ (instantiations) in those frameworks (Bonatti 1998). A fully specified theory (describing a complete set of inference rules or model-building procedures) can be refuted by empirical evidence, but a framework cannot. In a sense, the choice between frameworks is an ideological one. Ultimately, the strong argument from parsimony may tip the result; until mental logic can be shown to extend to non-deductive reasoning, its relevance to everyday life is open to question.
3.4.2 Content, Counterfactuals and Conceptual Structure

Humans rarely need to make deductions about abstract correlations between letters and numbers, or circles and triangles; in real life, the inferences we make are about real things and real situations. That is, they are contentful. As we have seen, realistic content can dramatically improve performance in the Selection Task, and it can affect the way in which conditional premises are interpreted.

Similar effects have been observed in ‘belief bias’ experiments, which demonstrate that the prior plausibility of a conclusion can aid or inhibit its acceptance by subjects. In one such experiment, Neuman (2003) asked high school students to judge a series of valid and invalid arguments for the existence of God; religious students were far more likely to accept invalid conclusions than were non-religious students. Goel & Dolan (2003) observed a belief bias effect in a syllogistic reasoning experiment where subjects were asked to judge the acceptability of belief-neutral and both ‘inhibitory’ and ‘facilitatory’ belief-laden syllogisms. Belief-neutral content was content where the subject would lack beliefs about at least one term in the syllogism. ‘Facilitatory’ content was content that was expected to facilitate a correct response, i.e. an implausible conclusion when the argument was invalid, and a plausible conclusion when the argument was valid. ‘Inhibitory’ content was the opposite, i.e. an implausible conclusion when the argument was valid, and a plausible conclusion when the argument was invalid. By way of illustration:

belief-neutral (invalid) :  
No codes are highly complex. 
Some quipu are highly complex.  
∴ No quipu are codes.

belief-laden/facilitatory (valid) :  
Some children are not Canadians. 
All children are people.  
∴ Some people are not Canadians.

belief-laden/inhibitory (invalid) :  
All calculators are machines. 
All computers are calculators.  
∴ Some computers are not machines.
Goel & Dolan perhaps unsurprisingly found that subjects performed better on the facilitatory tasks than on the belief-neutral tasks, and (slightly) better on the belief-neutral tasks than on the inhibitory tasks.

It therefore seems a necessary requirement of an adequate theory of reasoning that it explain the interaction between inference, representation and the general knowledge on which we base our judgements of plausibility. If not, then it will be a theory of little interest, having no explanatory power outside of the laboratory. This consideration is especially relevant for counterfactuals, which are more contentful and subject to the vagaries of belief and interpretation than almost any other kind of utterance. So I intend to address mental model theory's fifth objection to mental logic in a rather roundabout way. I shall not discuss the Selection Task or belief-bias experiments any further; rather, I shall investigate the analysis of realistic counterfactuals, and hope that this will throw some light on those other issues.

We recall the example of Goodman's match:

(13) If that match had been scratched, it would have lighted.

In conversation, we readily assent to (13). If someone tells us that the match was in fact scratched, we happily conclude that it lighted. But if we are then told that the match had been soaked in water overnight, we should conclude that the match did not light. We may even wonder why we were so quick to draw our initial conclusion without considering all the possibilities. But if we are also told that the match was of a special type that lights even when soaking wet, we will once more conclude that the match lighted. Even then, there will be infinitely many possibilities that we do not consider: what if the match is struck inside a bubble of nitrogen? What if we scratch the match on a surface that causes insufficient friction? What happens if God, unbeknownst to us, has decreed that matches do not light today? What if the match is a dud?

If pushed, we may protest that we are assuming a normal situation, and that it is impossible to consider all scenarios. Of course, what we really meant by (13) was something like (13'):

(13') If that match had been scratched, and the match was well made, and it was dry enough, and there was enough oxygen,..., it would have lighted.

This is fair enough. Whatever our cognitive capacity may be, it cannot be infinite, and we cannot consider an infinite set of possible scenarios. Humans
take shortcuts; in fact, humans take a lot of shortcuts, and take a lot for
granted. When we assert or assent to a counterfactual like (13), we assume
that the match was not soaked overnight, that there was enough oxygen, that
God has no particular interest in matches, and so on. Unless, of course, we
have some reason to believe that some inhibitory factors may obtain in the
referent situation. We may be underwater, or out in space; we may remember
a past experience where a match failed to light; we may believe in a rather
unorthodox reading of the Bible. In these cases, we may not be so hasty
to assent to (13); we will probably add the proviso *As long as we weren’t
underwater, then yes or suchlike.*

Byrne’s (1989) suppression of modus ponens operates in a similar way,
and would presumably work for Goodman’s match as well:

\[
\begin{align*}
\text{If that match had been scratched, it would have lighted.} \\
\text{If there had been sufficient oxygen, the match would have lighted.} \\
\text{The match was indeed scratched.} \\
\text{What, if anything, follows?}
\end{align*}
\]

I am certain that suppression would occur here, though the presuppositions
conveyed by the counterfactual premises could confound the issue by facil-
ilitating the conclusion *The match did not light,* on the basis that there was
not sufficient oxygen. On the other hand, I would predict no suppression for
the following scenario:

\[
\begin{align*}
\text{If that match had been scratched, it would have lighted.} \\
\text{If that match had been thrown in the fire, it would have lighted.} \\
\text{The match was indeed scratched.} \\
\text{What, if anything, follows?}
\end{align*}
\]

One should infer from this set of premises that the match did indeed light.
However, affirming the consequent and (to a lesser degree) denying the an-
tecedent inferences should be suppressed in this instance.

Mental model theory hopes to account for the problem of default assump-
tions via the Principle of Pragmatic Modulation. The initial model set for
(13) is of course

\[
\begin{align*}
\text{Fact: } & \neg \text{struck(match) } \neg \text{lights(match)} \\
\text{Counterfactual possibilities: } & \text{struck(match) } \text{lights(match)}
\end{align*}
\]
The fully fleshed-out models under a conditional interpretation should be as follows:

\[
\begin{align*}
\text{Fact:} & \quad \lnot \text{struck(match)} \quad \lnot \text{lights(match)} \\
\text{Counterfactual possibilities:} & \quad \text{struck(match)} \quad \text{lights(match)} \\
& \quad \lnot \text{struck(match)} \quad \text{lights(match)}
\end{align*}
\]

In the absence of additional information, we do not seem to arrive at this full set of explicit models; we do not consider the second counterfactual model in which the match is not struck, yet somehow still lights. We effectively arrive at a biconditional interpretation of the counterfactual. It is unclear from the literature whether this results from the effect of general knowledge on the fleshing-out process, or whether relevant general knowledge (or new information) causes the generation of a new set of models. However, it seems implicit in Johnson-Laird & Byrne (2002) and Byrne & Walsh (2002) that the latter is the favoured option.

Once a further necessary condition for lighting is made salient, however, we are forced to reinterpret our counterfactual. Doubts about the dryness of the match, whether raised by past experience, perceptual information or verbal information, will give us the revised models:

\[
\begin{align*}
\text{Fact:} & \quad \lnot \text{struck(match)} \quad \lnot \text{lights(match)} \\
\text{Counterfactual possibilities:} & \quad \text{dry(match)} \quad \text{struck(match)} \quad \text{lights(match)} \\
& \quad \lnot \text{dry(match)} \quad \text{struck(match)} \quad \lnot \text{lights(match)} \\
& \quad \ldots
\end{align*}
\]

Given this set of models, simply knowing (or hypothesising) that the match was struck is insufficient to allow the conclusion that the match lit (or would have lit). Similarly, our second example, the one where the match could also have been thrown into the fire, should yield the following set of models:

\[
\begin{align*}
\text{Fact:} & \quad \lnot \text{struck(match)} \quad \lnot \text{lights(match)} \\
\text{Counterfactual possibilities:} & \quad \text{struck(match)} \quad \text{lights(match)} \\
& \quad \text{fire(match)} \quad \text{lights(match)} \\
& \quad \ldots
\end{align*}
\]

This set of models poses no problem for modus ponens inferences, but it will (validly) suppress affirmation of the consequent and, should the models be made fully explicit, it will suppress denial of the antecedent as well.
A similar analysis can explain why contraposition, antecedent strengthening, and transitivity sometimes fail. Contraposition fails when the consequent of the premise cooccurs with the antecedent and the negation of the antecedent in different models, as in the ‘relevance’ interpretation of *If you’re interested in Vertigo, then it is on TV tonight*, or when the antecedent of the premise is not a necessary condition for the consequent. Antecedent strengthening fails when the strengthening proposition together with the premise antecedent triggers some piece of general knowledge or other information which puts the premise consequent in doubt. Transitivity fails when the antecedent of the first premise triggers some information that puts the consequent of the second premise in doubt, or vice versa. Chisholm’s Apollo sentences in (15) can also be explained: both sentences are true in different sets of models where different pieces of information have been triggered.

Of course, this is not much of an explanation at all. There is no account of why a given counterfactual or set of premises is interpreted the way it is, or how the relevance of diverse pieces of information is manifested. Once the appropriate mental models are in place, we can read off our conclusions with great ease, but the difficult task lies in accounting for the construction of the model set. In mental models theory, the line between interpretation and reasoning blurs. The partial account on offer is akin to a possible worlds analysis without similarity, and lacks predictive power.

Mental logic faces an equivalent problem, that of identifying the appropriate set of relevant beliefs with which to evaluate a particular counterfactual. We recall that the metalinguistic analyses of counterfactuals stumbled at the same place. Yet the mental logicists face a further problem, that of describing the interaction between representation and the autonomous reasoning mechanism. This problem is not shared by mental models, which merges the two; as soon as a piece of information is available in a model, it is available for reasoning with. If, halfway through a proof, it becomes necessary to evaluate the conditional *If the match had been struck, it would have lighted*, where does the ‘relevant information’ come from? It seems unlikely that it is pre-loaded as a set of premises when the reasoning routine is initialised, as it may not be possible to predict which conditionals will need to be evaluated. However, if we accept that representation and reasoning interact during the deductive process, then the lines between the two start to blur and the idea of formal rules begins to lose its integrity. Braine & O’Brien (1991) accept that extralogical factors can influence or even bypass the logical reasoning process, but they are quiet on the subject of whether such factors affect the
process while it is running. I cannot answer on their account; here I will concentrate on the first issue, the problem of relevant beliefs.

Johnson-Laird (2001) recognises that this is a significant problem for the model theory in general. Yet, as Johnson-Laird & Byrne (2002) observe, nobody knows how knowledge is represented in the mind. Braine & O’Brien (1991), faced with the same problem, state that “the solution to this issue must wait for a full account of reasoning” (p. 239). A full account of reasoning would also be a full account of mental representation. A few contenders have been proposed, notably propositional representations in a mental language (Fodor 1975), parallel distributed processing (Rumelhart & McClelland 1986) and the SOAR framework (Newell 1990). However, none of these models has been developed sufficiently to capture the extremely complex problem at hand. We shall nonetheless consider some aspects of how knowledge may or may not be encoded in the mind, as the question is one of great importance for our understanding of counterfactual thought.

Firstly, it is possible to say a few things about how the search for relevant conditions cannot be implemented. It cannot be the case that we sequentially examine each of our beliefs and check whether it, either directly or taken together with other beliefs, supports or rules out a conclusion. This tactic is equivalent in computational complexity to checking a knowledge base for consistency, which is known to be highly intractable for any knowledge base of non-trivial size. Such problems are NP-complete, meaning that it is very unlikely that they can be solved by any polynomial time algorithm, i.e. that they are computationally intractable. NP-complete problems, such as the famous Travelling Salesman problem, have received a great deal of attention in computer science, due to their usefulness, their prohibitive complexity and not least to the fact that all NP-complete problems reduce to the satisfiability problem in propositional logic – the problem of finding an assignment of truth values that satisfies a set of logical clauses. If a polynomial-time algorithm were to be found for one NP-complete problem, all NP-complete problems

---

20That all NP-complete problems can be reduced to the satisfiability problem was proved by Cook (1971). A wide range of important problems have subsequently been shown to be NP-complete, including: finding a Hamiltonian circuit in a graph; checking a relational database for violations of Boyce-Codd Normal Form; ‘instant insanity’; checking whether a formula of propositional logic is non-tautologous; checking non-equivalence of finite state automata; membership of an indexed language; checking graph isomorphism; and, of course, the Travelling Salesman problem (see Garey & Johnson 1979 for an overview of NP-completeness and related issues).
would necessarily be tractable. However, most computer scientists conjecture that no such algorithm exists. Assuming the prevalent ‘computational brain’ hypothesis that the human brain is a computer in the broad sense of the term, i.e. that its functionality can be replicated by some Turing machine, it cannot be expected to reliably solve any intractable problem. The fact that the brain is capable of extremely fast performance on some tasks does not make intractable problems any more tractable; increasing the processing speed of a computer 1,000-fold merely adds 10 to the size of the largest instance of an order $2^n$ algorithm that can be processed in an hour (Garey & Johnson 1979). Given the amount of information stored in each person’s brain, it is inconceivable that we search through all of it every time we consider a counterfactual.

So we must be a little less ambitious; but no matter. Humans are not ideal reasoners who have access to the deductive closure of their beliefs. After all, when we think about a match lighting, we can (and do) happily ignore information relating to yesterday’s weather, religion or the works of Flann O’Brien. Johnson-Laird & Byrne (2002) assume that long-term memory represents knowledge as explicit mental models, and that this knowledge can modulate models in working memory. They describe a computer program that implements the Principle of Pragmatic Modulation, as applied to Goodman’s match. The program utilises a knowledge base containing sets of models that correspond to rules about the world. The following model represents the rule If a match is soaked, it will not light (p. 659):

\[
\begin{align*}
&\text{soaked}(\text{match}) \quad \neg \text{lights}(\text{match}) \\
&\neg \text{soaked}(\text{match}) \quad \neg \text{lights}(\text{match}) \\
&\neg \text{soaked}(\text{match}) \quad \text{lights}(\text{match})
\end{align*}
\]

Learning that the match was soaked will cause the first model in the rule set to be integrated into the set of models of the premises, and lead to the conclusion that the match did not light$^{21}$. Activating the second model for

$^{21}$Under the standard method of integration, this model would contradict the initial set of models for If a match is struck, then it lights. However, Johnson-Laird & Byrne state that models from general knowledge take precedence when they contradict premise models, and can simply overwrite them. One suggests that a more sophisticated version would invoke differing degrees of belief. Knowledge about soaking matches never lighting might possess a high degree of belief, but it could be overruled by actual perception of a soaking match lighting. Providing a full theory of belief weighting is of course a daunting task, and probably also requires a “full account of reasoning.”
the rule allows us to generate the counterfactual *(If the match had not been soaked and been struck, then it might have lighted).* Of course, what most people would say is *(If the match had not been soaked and been struck, then it would have lighted)*, i.e. they would not consider models where the match was dry but still did not light, unless some additional factor was made salient. If we are to consider this program as a model for reasoning in general (which the authors have not suggested), then we note a number of flaws. The rules are highly specific and do not generalise to other cases. It is not indicated that if a match is soaked in petrol, then it still generally lights. We would need another rule for that case, or at least one for inflammable liquids in general; we would also need rules stating that other things light when they are not soaked. We would need to specify what happens when two rules in general knowledge contradict each other, such as when a second rule states that ‘magic matches’ light even when soaked. On the other hand, this approach is more realistic than the first one we considered, as it only consults a restricted set of knowledge. If we confine our search to information contained in the current set of models, then we can efficiently trawl our knowledge base for relevant information. But then our task is to describe how new information comes into our mental models unprompted, whether by association or analogy or by some mystical creative process.

In computer science, search algorithms are often made easier by storing data in suitable data structures, e.g. binary search trees. Such structures implement some ordering on the data, so that the algorithm can quickly hone in on the part of the structure where the target of the search is guaranteed to be found. We might entertain the idea that information in the brain is structured in a similar way. Our cognitive search strategy should concentrate on information relating to match striking, which might inherit features from knowledge about fire in general, including the necessity of oxygen and dryness. In this way we might identify additional information that must be considered, but which does not explicitly share content with our set of models. However, this suggestion leads to significant problems of redundancy and generalisation. Oxygen is necessary for fire, but it is also necessary for breathing and for oxidisation; need we represent the information that oxygen is generally present on Earth in every part of the data structure where oxygen is necessary? We could assume that every structure inherits basic knowledge about default environmental conditions; it then becomes problematic to constrain these conditions. It is impossible to specify all the factors that could possibly affect the lighting of a match – yesterday’s weather becomes relevant.
if the match may have been left out in the rain; as we have already considered, God may decide to intervene\textsuperscript{22}. We must ultimately conclude that each item of information is connected to every other item of information. This conclusion may not be overly problematic for a general theory of knowledge representation, but it does not bring us any closer to identifying how beliefs become relevant to a counterfactual.

In any case, selection of thematically related beliefs may not be the only cognitive process underlying the representation of counterfactuals. Yarlett & Ramscar (2001) identified similarities between counterfactual and analogical reasoning. In an experiment, they gave subjects a story about unfamiliar material and either a story with similar surface features or a story with similar relational structure. The subjects were then asked to rate the plausibility of a variety of counterfactual scenarios based on the main story. The responses of subjects who had read the surface-similarity story did not differ significantly from those recorded in a pre-test where only the main story was used; subjects who had read the relational-similarity story showed an increase in the plausibility of counterfactuals that corresponded to that story. In a second experiment, a distraction task was introduced between the presentation of the similar story and the presentation of the main story. The behaviour of the surface-similarity group was no different, but the facilitatory effect of the relational-similarity story disappeared. Yarlett & Ramscar note that analogical matches also tend to be difficult to retrieve from memory, and reach the conclusion that relational similarity can in certain cases play an important role in counterfactual thinking, just as it does in analogical reasoning.

Like Lewis and Stalnaker, we must turn our attention to the matter of similarity. Judgements of similarity have been more studied in cognitive psychology than even counterfactuals have, and it would be impossible to attempt a satisfactory account of the field here. When we consider the possibility of a match lighting, we may think about previous instances where a match was struck, and see whether the match lit in those instances. If the match lit in some instances but not in others, we may consider which

\textsuperscript{22}I am not sure how the work of Flann O’Brien could realistically influence whether a match lights or not. Nevertheless, artificial examples can be constructed, e.g. an unusual game where successfully identifying the authors of famous books determines whether participants get to light a dry or a wet match: Participant A claims that Flann O’Brien wrote \textit{Syntactic Structures}; Participant B does not see A’s match, but knows that his answer was incorrect, and that he has been given a wet match: A does not light his match, but B confidently asserts \textit{If A had struck that match, it would not have lighted}. 
instances were more similar to the present one. An influential account due to Tversky (1977) proposes that the process similarity judgement involves an overall comparison of two objects on the basis of their shared and distinct features. Hahn et al. (2003) provide a different model, based on the concept of *Representational Distortion*, or the complexity of the algorithm required to transform one object into the other. This theory has the advantage that, unlike that of Tversky, it can judge the similarity of structured representations. For example the sequence of numbers 1 2 3 4 5 is judged to be more similar to the sequence 11 12 13 14 15 than to the sequence 1 2 3 0.5 9, as only one kind of operation (add/subtract 10) is required for the former transformation, whereas two kinds of operation are required for the latter.

It is conceivable that a similar process may decide the degree of similarity between a counterfactual situation and actuality.

It is sometimes the case that we consciously consider the similarity of counterfactual scenarios, but more often they just ‘leap out’ at us. If similarity in surface or relational structure does underlie our use of knowledge in counterfactual reasoning, then it must also do so at an unconscious level. In order to investigate the properties of the process by which particular situations or models become more available than others, we must investigate the properties of the kind of things that tend to just ‘leap out’. Tversky & Kahneman (1982) described an ‘availability heuristic’ that is used to make rapid judgements of frequency or probability: “A person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind” (p. 164). They found that use of this heuristic can lead to systematic biases. For example, one experiment showed the effects of familiarity on frequency judgements: subjects were read 2 lists of names, consisting of either a list of 19 famous men and 20 less famous women or 19 famous women and 20 less famous men; they were then asked whether there were more men or women listed; 80% of subjects thought that the list of famous names was larger than the list of less famous names. Tversky & Kahneman attributed this bias to reliance on the availability heuristic. In general, they hypothesised that dramatic and salient (or otherwise “memorable”) situations can fool the heuristic into assigning them high subjective frequency. When we imagine alternate scenarios, we tend to concentrate on the ones that are for one reason or the other, most available: “If many [such] scenarios come to mind, or if the one scenario that is constructed is particularly compelling, the event in question appears probable” (p. 177). This implies the existence
of a more multi-faceted procedure than an objective similarity metric, most likely involving degrees of belief or probability weightings. But this should come as no surprise, for our counterfactual thoughts are open to many kinds of influences: they may be goal-driven ("I wonder how things could have turned out better"), they may be affected by constant preoccupation with a certain scenario, or by emotions.

One approach that appears promising is connectionism, or parallel distributed processing (Rumelhart & McClelland 1986). The connectionist memory of memory (presented by McClelland & Rumelhart 1985) is different from the conventional view provided by our experience of computers; it holds that memory is massively distributed across a large network of interconnected processing nodes, each of which represents a 'microfeature'. An input to the network (whether perceptual or from another cognitive process) will cause a certain pattern of activation corresponding to the microfeatures it possesses. Moreover, that pattern will strengthen the connections between the nodes that have activated, and weaken the connections between those that have not. Learning thus merely involves altering connection weights. What we consider to be a 'discrete' memory is not stored in any one location, but as a pattern that overlaps with many other memories. One of the strengths of the distributed model is its explanation of abstraction. For example, we learn the concept of 'dog' by observing many different dogs in different situations. We arrive at an abstract concept of 'dog' that is not identical to any individual dog we have observed, but which shares many properties with all of them. We can also explain similarity in terms of patterns of activation. In general, a dog will be judged more similar to a cat than to a bagel as the pattern activated by 'dog' shares more microfeatures with the pattern activated by 'cat' than with the pattern activated by 'bagel'. The network can implement Wittgensteinian family resemblances by identifying an overlapping tendency across individual patterns of activation, so that there is no pattern corresponding directly to 'game', merely patterns corresponding to types of games, each of which is sufficiently similar to another type of game.

How the distributed memory network responds to an input pattern will depend on its relation to patterns of activation that have been stored in memory (cf. Rumelhart 1989):

1. When a previously stored pattern is input, the network reacts by amplifying the pattern of activation and strengthening the connections between the activated units.
2. When an unfamiliar pattern enters the network, the system responds weakly or not at all.

3. When part of a previously stored pattern enters the network, the network completes the pattern.

4. When a pattern is input that resembles a previously stored pattern, the input is distorted towards the familiar pattern.

Our construal of a given situation is thus dependent on our experience of similar situations. When we think about striking a match, the corresponding pattern of activation will be modulated by previous match-related inputs. If the weightings in the system make the presence of oxygen salient for match-lighting (because our attention has been drawn to this condition by past experience, or because someone told us a story about lighting fires on the Moon), then the current situation will activate the requirement for oxygen and it will suddenly ‘leap out’ at us that oxygen must be present. The pattern for the oxygen condition will be shared by all fire-related patterns, and by breathing and oxidisation patterns. We thus resolve the problem of redundancy, and come closer to an understanding of how relevant conditions might be identified. Implementing the availability heuristic will require that more ‘memorable’ situations lead to a more substantial change in weightings than do more normal situations. Accounting for ‘memorability’ is another task.

There are many benefits to the connectionist model. Unlike other paradigms in cognitive science, it is ‘biologically inspired’; the interaction between units in a network is modelled on the interaction between neurons in the brain. Connectionism supplies a general theory of memory and learning. A connectionist approach to reasoning would integrate smoothly with accounts of other areas of cognition, whereas formal approaches have been successful in illuminating only a few domains, such as deduction and language. Parallel processing brings with it the advantage of great speed. However, there are also problems with the theory. The connectionist enterprise is still in its infancy; while neural networks have proven very successful at simulating

---

23Oaksford & Chater (1998: 38), arguing for a connectionist approach to reasoning, claim that “The only experimental work we know of that explicitly addresses the logical characteristics of the cognitive architecture is that on deductive reasoning.” They thus seem to ignore the vast amount of work in psycholinguistics that is informed by very logical theories of language.
behaviour in some areas, e.g. low-level vision, detailed accounts of higher cognitive functions remain beyond our grasp at present. Certain crucial logical processes, such as variable binding and recursion, which seem integral to any account of language (at least), are extremely difficult to model in connectionist networks (Ellis & Humphreys 1999). Fodor & Pylyshyn (1988) have argued that the connectionist model of distributed representation cannot explain structural relationships such as that between *John loves Jane* and *Jane loves John*. It is therefore unclear that neural networks can adequately capture the analogical process, which is insensitive to bundles of unstructured features (unless, of course, structure is somehow encoded in microfeatures). In contrast, traditional logical frameworks have been quite successful at describing many aspects of analogical reasoning (e.g. Keane 1997).

Interestingly, Rumelhart (1989) describes how mental models might be implemented in a connectionist network, although he claims that similarity plays a greater role than models (or formal rules) in most everyday reasoning. A mental model corresponds to a network that reacts to specifications of actions (e.g. premises) and relaxes to a state that integrates these specifications. The network can then feed its ‘conclusion’ (new state) back into the memory network as perceptual or conceptual input, depending on the task at hand.

This is essentially a two-process model of reasoning\(^{24}\) that distinguishes between reasoning by similarity (emergent properties of the memory system) and reasoning by simulation (mental models). It is unclear where the division of labour between the two processes lies; Rumelhart seems to identify model-based reasoning with visual reasoning, which is clearly a narrower conception that that of Johnson-Laird and Byrne. Yet the general framework is not incompatible with the mental model theory. A distributed memory system may be responsible for things we ‘just know’ without any active consideration, and hence for strong content or belief bias effects that bypass the reasoning system entirely. Actual reasoning, involving the investigation of alternative possibilities, is done by a network, corresponding to a mental model, that receives input from the memory system (relevant information) and draws a

\(^{24}\)Rumelhart recognises that we may also reason using formal rules we have learned. However, he assumes that reasoning with formal rules involves simulating graphical procedures (‘manipulating the external world’) in a mental model. When we do complex mental arithmetic, we use the same algorithm as we would on paper; we simply visualise the individual steps in a model.
conclusion. This system is also susceptible to content effects, as the input from memory will be modulated by the comparative availability of alternative scenarios.

Such a division is in fact implicit in the model theory. It is impossible to store all our knowledge in working memory, but it is only in working memory that reasoning occurs. If we adopt Johnson-Laird’s (1983) description of long-term memory as a ‘model of the world’, then we arrive at a system with two levels of models. Evans (1993b) has suggested a similar two-process model of reasoning, in which a connectionist network explains ‘implicit’, familiar-content reasoning and mental models are utilised for ‘explicit’ cognitive tasks involving deduction and mental simulation. There is some neurophysiological evidence for a heterogeneous reasoning mechanism. Goel & Dolan’s (2003) fMRI studies of belief modulation in syllogistic reasoning identified two distinct areas of brain activation: belief-laden problems tended to activate a left temporal system associated with semantic retrieval, while belief-neutral problems activated a bilateral parietal system associated with the internal representation and manipulation of spatial information. Interestingly, this latter brain area has also been implicated in certain kinds of arithmetic reasoning that involve the approximation of quantities, as opposed to symbolic manipulation or the use of tables that have been learned by rote (e.g. Dehaene et al. 1999, Cohen et al. 2000). Dehaene (1992) describes how tasks such as spontaneous number comparison or magnitude estimation “solicit an ‘approximation mode’ in which we access and manipulate a mental model of approximate quantities similar to a mental ‘number line’” (p. 20; italics added). It is not clear whether the term ‘mental model’ is used here in any technical way, but the congruence is nonetheless rather interesting. Knauff et al. (2002) studied patterns of brain activation in conditional reasoning tasks (modus ponens and modus tollens inferences from belief-neutral material), and also identified activation in a bilateral occipito-parietal network. These authors concluded that “the obtained activation...provide evidence that reasoning is a cognitive process in which spatially organized mental models are used for reasoning” (p. 209), and held that their results contradicted the mental logic position that reasoning involves symbolic manipulation mechanisms located in brain areas associated with language.

However, the neurophysiological evidence is far from conclusive. A literature review by Wharton & Grafman (1998) indicates the diversity of results found in the cognitive neuroscience literature. These authors suggest that the balance of the evidence supports left-hemisphere activation,
particularly in the language-processing areas, for content-independent rea-
soning tasks; they note that aphasics and left-hemisphere-lesioned subjects
have difficulty in solving simple relational deduction problems, whereas right-
hemisphere-lesioned subjects do not. They implicate right-hemisphere sys-
tems in content-dependent reasoning. Wharton & Grafman’s main conclu-
sion is that experimental evidence does not support mental model theory’s
prediction that visuospatial processing areas are involved in all kinds of rea-
soning (Johnson-Laird 1995); nor is this conclusion compatible with the ev-
idence provided by Goel & Dolan or Knauff et al. One might suggest that
neuroimaging of reasoning is bound to be an imprecise task, given the un-
avoidable complexity of the stimulus and the number of confounding variables
that may affect the analysis.

No full account of content and belief effects exists at this point in time,
and it will be many years before one becomes available. However, some
preemptive remarks may be made. In contrast to mental logic, mental model
theory has the significant advantage that it can at least describe such effects
in terms of the comparative availability of models – more realistic models are
simply more available than less realistic ones. Mental logic faces the problem
of accounting for content effects on the formal reasoning process which, unlike
the mode of representation, is held to be independent of content. We have
briefly considered connectionism as a third account of reasoning, at least for
the intuitive content-rich inferences that we make every day. The strength of
the connectionist account lies in its integration with the memory system and
its powerful pattern-matching machinery, and in the resultant fact that the
reasoning system can accommodate much more ‘relevant information’ than
the relatively constrained mental models reasoning process. We have also
considered the possibility of a two-process reasoning system, which might
involve a mental model reasoning system superimposed on a connectionist
distributed memory system.

3.4.3 Conclusion

Psychological theories of counterfactuals have for practical reasons tended
to concentrate on their interpretation and representation, and have had lit-
tle to say about their generation. One exception is Byrne & Tasso’s (1999)
study of abstract counterfactual generation, which demonstrated that sub-
jects assert counterfactuals based on the contents of their mental models.
Of course, we have no account of how mental models are populated in non-
CHAPTER 3. PSYCHOLOGICAL PERSPECTIVES

experimental conditions, and we revisit the problem of relevant information. This core problem, which faces all psychological explanations of counterfactuals, is analogous to that facing the philosophical theories we considered in chapter 2. The function of ‘availability’ is not unlike that of ‘similarity’: just as philosophers have failed to identify how possible worlds are ordered, psychologists have failed to identify how one situation can seem more available than another. This is made clear by the explicit congruence between Braine & O’Brien’s (1991) Schema for Conditional Proof and the metalinguistic theories of Goodman and Chisholm. Just as the latter’s attempts to arrive at a satisfactory definition of cotenability foundered without reference to possible states of affairs, mental logic cannot account for ‘relevant information’ without situating it in an appropriate theory of representation. In general, mental model theory has been more successfully developed than mental logic, but we have seen that the latter theory is not as explanatorily deficient as it is sometimes made out to be.

All of the theories that have been considered in this chapter reflect the prevalent cognitivist paradigm in psychology, which equates cognition with computation. Cognitive systems are treated as ‘black boxes’, defined in terms of their output relative to input. An adequate theory of reasoning is thus one that can accurately simulate and predict human behaviour on reasoning tasks. Cognitivist theories are independent of implementation, in that they hold functionality to be unconstrained by the nature of the ‘hardware’ – a program running on a computer is equivalent to a program running on a brain if and only if they exhibit the same functionality. The theory of mental logic most explicitly captures this notion of cognition as symbolic manipulation; however, Johnson-Laird (1983) stresses the computational characteristics of mental models, and the theory has been implemented in numerous computer programs (e.g. Bara et al. 2001, ní Bhriain 2002). In spite of its claims to be ‘biologically inspired’, most researchers in connectionism also adopt a high-level approach. A number of integral mechanisms in the connectionist analysis, such as back-propagation networks, have no known biological correlate (Ellis & Humphreys 1999). As Rumelhart (1989) puts it, the distinctive feature of connectionism is not that it eschews computationally (which it does not), but rather that views most cognitive psychology as based on “the wrong kind of computer” (p. 299).

The cognitivist enterprise has been very successful at illuminating many areas of psychology, but its limitations may eventually be exposed, just as those of behaviourism were before it. There is still no high-level account of
emotions, despite their crucial role in driving reasoning – counterfactuals, in particular, are triggered by, and ultimately effect, our emotional processes. Why do we try to imagine a different past, when such a past is clearly impossible? Why do we need to know that things could have turned out differently? The reasons for counterfactual thought may ultimately prove more mysterious than the processes underlying it. Haugeland (1981) offers a critique of cognitivism on such grounds:

What is common sense, let alone creativity, wit, or good taste? What happens when we fall asleep, let alone fall under a spell, fall apart, or fall in love? What are personality and character, let alone identity crises, schizophrenia, the experience of enlightenment, or moral integrity? We turn to psychology if we think these questions have scientific answers; and if we shouldn’t, why shouldn’t we? Cognitivists are as vague and impressionistic on such issues as psychological theorists have always been.

(Haugeland 1981: 225)
Chapter 4

On The Relationship Between Philosophy and Psychology
4.1 On Logic and Thought

“This idea seems so contrary to common sense that, as you might suspect, it has been advocated by philosophers (and psychologists).”

“To sum up briefly, it is the business of the logician to conduct an unceasing struggle against psychology and those parts of language and grammar which fail to give untrammelled expression to what is logical.”
– Frege (1979: 6–7)

Psychology, like so many other disciplines, came into being as an offshoot of philosophy. Yet the filial relationship has often been a fraught one, not least in matters of logic and reasoning. Frege, the father of modern logic, attacked what he saw as the wrongheaded conflation of psychology and logic espoused by psychologist philosophers such as Husserl. The title of Boole’s (1854) influential work, An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities represents a common psychologist position, that logic is a product of the human mind. This position was anathema to Frege, who held that psychology and logic had little in common, and that psychology had nothing at all to say about truth: “In order to avoid any misunderstanding or any blurring of the border with psychology, I attribute to logic the task of finding the laws of truth, not those of perception or thought” (Frege 1918; my translation).

Frege defended truth as a metaphysical notion independent of our grasp of it. If logic resides in our minds, he argued, then how can people make mistakes? How can anything have been true before humans existed? Many truths are unknowable, he argued, but they are no less true for it. Our knowledge that something is the case does not cause it to be true; the truth is the truth, whether or not we know it.

The theme of this chapter will be the relationship between philosophical and psychological research on counterfactuals. It might be suggested that this smacks of psychologism; however, I do not propose to concern myself with absolute truths, but rather with the nature of the truth judgements that

\(^1\)To be precise, Frege attacked Husserl’s early work. Husserl later abandoned psychologism and adopted the theory of phenomenology for which he is famous.
humans commonly make about counterfactual statements. As I previously
mentioned, I am not ‘doing metaphysics’. It is probably impossible for us
to know how the world would be if Apollo were human, or what would have
happened if the CIA had assassinated Hitler in 1937. There may well be
objective truths about these matters, but even if they are, they are beyond
our grasp. Maybe we should always respond to counterfactuals with silence,
like the wise Stoic. In practice, however, we are rarely happy to suspend
judgement when someone suggests a counterfactual with which we do or do
not agree. But on what basis do we agree (or not)? Many philosophers who
have studied counterfactuals have done so from a metaphysicist’s perspective,
yet that is not to say that their theories offer no insight into the thoughts of
mortals who are neither omniscient nor infallible. The opposite is in fact the
case.

4.2 On Bounded Rationality

Philosophers often address epistemological problems from the viewpoint of
an ideal reasoner. An ideal reasoner would find modus tollens inferences
unproblematic, and easily succeed at the Selection Task. An ideal reasoner
would not believe that dandelions are oracles of love, that Friday the 13th
is always unlucky or that alligators live in the New York sewer system². An
ideal reasoner would not play blackjack. An ideal reasoner would always
win at chess, unless he was playing another ideal reasoner, in which case
there would be a stalemate. An ideal reasoner would have devised a far
superior account of counterfactuals than has yet appeared. Humans, it seems
clear, are far from ideal reasoners – we systematically fail simple reasoning
problems; we are superstitious; we gamble when the odds are poor; games of
chess usually have winners; the problem of counterfactuals remains unsolved.

When we think about striking a match, we cannot consider every possible
factor that might affect the outcome. Firstly, it is self-evident that we cannot
consider things that we do not know about, and we are not omniscient. It
may be that some as yet undiscovered physical property of melons inhibits
the lighting of matches in close proximity to them, but if this is the case, then
I am ignorant of it. I will confidently assert that if I had struck my match it
would have lighted, even when there is a melon nearby. In Ramsey’s terms,
the effect of melons on matches does not belong to my stock of beliefs. In

²Unless, of course, he actually saw one.
any case, it would be unfair to criticise humans for their non-omniscience, as we have never claimed to be otherwise.

There is a second sense in which we can be considered less than ideal reasoners. Consider the counterfactual in (31):

(31) If gravity didn’t exist, then you and I could fly around.

To me, this seems like a reasonable thing to say. I would probably assent to it. However, it is almost certainly as false as any counterfactual can be. It is hard to imagine how a universe without gravity would actually be, but it is highly probable that there would be no planet Earth, and no humans. ‘You’ and ‘I’ wouldn’t be in any position to fly around, as we wouldn’t even exist. However, I think that (31) is acceptable. One might argue that the speaker is not really envisaging a situation where there is no gravitational force in the universe, merely one where some localised temporary suspension of gravity obtains. I do not believe that this is the case; the counterfactual means what it says, namely how things would be if gravity simply did not exist. Yet if (31) is blatantly false, and I know it is false, how could I judge it true? The reason is that when we add the hypothetical non-existence of gravity to our stock of beliefs, we do not consider all its consequences. In a sense, we merely stipulate that gravity doesn’t exist, and fit the actual world around that stipulation without investigating the repercussions. If I am a physicist, or a pedant, or if the effects of gravity are particularly salient to me, then I may very well object to the counterfactual, but in general, I find it conversationally unproblematic. A (slightly) less lurid example of the same phenomenon is provided by (32):

(32) If I had been alive in 1937, I would have assassinated Hitler.

The world would have to be very different if I were to have been alive in 1937, whether I travelled there in a time machine, or was born 60 years earlier than in the actual world. Hitler might not have been around in the closest such possible world; one cannot say. Yet the meaning of the counterfactual is simple – it positions me in the year 1937, and leaves everything else as it is and was.

Stalnaker (1968) writes that “the rational man accepts the consequences of his beliefs” (p. 101). This is exactly what the sentences in (31) and (32) require us not to do. Is man then irrational? Before answering this question, we should consider the nature of rationality. We can be said to be rational if our reasoning behaviour is in accordance with some normative system, usually
logic. Subjects’ performance in the Selection Task thus appears irrational. It is harder to establish a normative criterion for counterfactuals, but it is far less certain that a metalinguistic or possible-worlds theory licenses (31) and (32) than it is that they are conversationally acceptable. Humans possess extremely limited resources for limited deduction, and frequently make errors; in a strict sense, we are irrational animals.

The ‘paradox of the preface’ (Priest 1998) offers another example of seemingly ‘irrational’ reasoning. A hypothetical author writes a non-fictional book whose claims are based on evidence $\alpha_1, \ldots, \alpha_n$. This evidence is as convincing as is empirically possible, so the author rationally endorses $\alpha_1, \ldots, \alpha_n$. However, the author is aware that virtually all substantial factual works contain at least some false claims; he therefore rationally endorses $\neg(\alpha_1, \ldots, \alpha_n)$. The author does not believe that $\alpha_1, \ldots, \alpha_n \land \neg(\alpha_1, \ldots, \alpha_n)$. He has behaved rationally, yet he rejects the consequences of his beliefs. How can the author be said to be rational?

*Bounded rationality* refers to rational behaviour subject to cognitive processing constraints (Evans 1993a). It operates on the principle that it is unfair to criticise someone for not doing what they cannot possibly do. We can say that humans are rational in the bounded sense if they have the basic ability to make logical inferences, but are hampered by the limitations of memory and processing speed. According to the two psychological theories we considered in chapter 3, this is indeed the case. Mental logic is explicitly based on rational logical principles – we make errors when we lose track of a proof, or accept something we believe to be true without engaging our reasoning processes. As we have seen, mental model theory is also inherently logical. A computer that reasoned by manipulating mental models would be very rational, although its rationality would at some point be bounded by memory limitations. The advantage of mental models over mental logic is that it preserves bounded rationality while more effectively facilitating content and belief bias effects. Hence Johnson-Laird & Byrne’s (1993b) remark that humans “are rational in principle, but err in practice” (p. 205).

A similar distinction is found in the study of language. After Chomsky (1965), it is common to speak of linguistic competence, or linguistic knowledge in the mind, and linguistic performance, or language as it appears in the world. The generative power of humans’ internalised grammars is such that they can generate infinitely long sentences; for example, any number of shorter sentences can be conjoined to generate utterances like *The weather is fine and it is summer and the birds are singing and there is milk in the
Humans thus have the capacity to create sentences that they could never possibly say. Our mental grammar is theoretically capable of generating a sentence that would take 200 years to produce, but we would die before we got to the end of it. More importantly, our ability to produce and comprehend linguistic utterances is constrained by working memory limitations. The sentence in (33a) is not significantly different in structure to (33b) or (33c), but (33b) is extremely difficult to parse, and (33c) is nigh on unparsable for any human.

(33)  
a. The man the dog bit screamed.
b. The man the dog the penguin attacked bit screamed.
c. The man the dog the penguin the cat chased attacked bit screamed.

Our linguistic competence is ‘rational’, in that sentences with multiple nested reduced relative clauses are in theory generable. However, our cognitive limitations place a bound on the externalisation of competence, i.e. on performance. Likewise, it is suggested, with mental logic or mental models.

Given bounded rationality, can it truly be said that human deductive reasoning is deductive in nature? If a deductive conclusion is drawn from a set of premises, then there is no way that the premises can all be true and the conclusion false. Clearly, Johnson-Laird & Savary’s (1999) illusory inferences represent a complete failure of deduction, resulting from non-consideration of every way in which the premises can be true. When we reason about matters in the world, we cannot consider every possible fact that might influence events. We must base our conclusions on deductively insufficient premises. When we consider a match being struck, we automatically infer that the match lights. If the match happens to be soaking wet, then we must concede that our premises were true, but our conclusion false. ‘Common sense’ reasoning is therefore nonmonotonic: new information can render our previous conclusions false. A classic example of nonmonotonic reasoning relates to default or generic rules: Birds fly. Tweety is a bird. Therefore, Tweety flies.

This argument appears valid; if, however, we later find out that Tweety is an ostrich, we will revise our belief in the conclusion, not in the premises. When we reason with counterfactuals, as with generics, we are making inferences that are inherently defeasible. The development of nonmonotonic logics in recent years raises the possibility that human rationality may be saved, given an appropriate normative criterion. However, even a logic of induction (or some other kind of nonmonotonic inference) would be subject to cognitive limitations – the task of checking for validity in classical logic
is NP-complete, and validity checking in nonmonotonic logics is even more complex (Oaksford & Chater 1993).

Mental logic and mental model theory both claim to avoid the pitfalls of intractability. Braine & O’Brien (1991) state that consistency checking is not a part of the mental logic reasoning system. They observe that people may hold contradictory beliefs (as illustrated by the paradox of the preface), without ever arriving at a reductio ad absurdum, and they suggest that validity is not a concern for human reasoners unless their conclusions are explicitly challenged, either by the actual world or by a conversational partner. So long as we never have to reason from two contradictory premises, we will not even be aware that a contradiction obtains. If, as seems a minimum necessity, mental logic is as expressive as the predicate calculus, then the spectre of undecidability must be faced – that there is no finite method of determining whether an argument is valid or not. In particular, there will be infinitely many possible derivations from a given set of premises, and if an argument is invalid, it is impossible to show this by exhausting all possible proofs. Mental logic seeks to curtail infinite derivations by limiting the use of recursive ‘feeder schema’ in derivations. Furthermore, it may simply be that performance-related factors prevent a deduction from proceeding ad infinitum, so that after a certain number of steps the reasoning routine just ‘gives up’ and frees up processing resources.

Mental model theory enlists a number of mechanisms in its attempts to escape intractability. Firstly, a given mental model only represents a (very) partial subset of the world. Secondly, the fleshing-out process is subject to the limitations of working-memory, and frequently fails to exhaust all possible models. When we reason about everyday matters, however, we must keep many pieces of information in mind, even for a simple task such as cooking pasta. In such scenarios, given the acute processing constraints that the theory posits in general, it might appear that the mental models in working memory would become overloaded. There are two ways to avoid this problem: (i) it may be the case that representing a large amount of facts in a single model is much easier than representing multiple (even 2 or 3) models that contain few facts (this conception of everyday reasoning as seeking ‘truth in a model’ is certainly appealing); or (ii) a ‘two-process’ system of reasoning such as was considered in section 3.4.2 could offer a division of labour among two subsystems whose strengths lie in different reasoning domains.
4.3 On Modes of Representation

In the previous chapter, the role of representation in reasoning was stressed. Almost no contexts apart from psychological experiments, games, and research involving logical or programming languages require that we make inferences about abstract entities that do not relate to the world in any way. General knowledge must be recruited in order to consider any contentful counterfactual, and it is crucial for any account of counterfactuals that it explain the representation of salient knowledge. Philosophical constructs such as possible worlds and situations also constitute modes of representation, and it will be useful to consider how these compare with Johnson-Laird and Byrne’s mental models. A direct comparison with mental logic is not possible, as it does not offer a means of representing possible states of affairs. One guiding consideration in the following discussion will be the importance of partial representation – the principle of bounded rationality imposes on us the requirement that we cannot represent everything that there is.

The exact nature of possible worlds has been the subject of much discussion among philosophers, and there is no real consensus on the issue\(^3\). David Lewis has defended a position of ‘modal realism’ that holds that possible worlds are no different in nature from the actual world, whose sole claim to preeminence is that it is the one that we happen to inhabit. Possible worlds have an existence independent of our universe, and are not connected to it in any way. This thesis permeates most of Lewis’ works; the following quote is illustrative:

> I emphatically do not identify possible worlds in any way with respectable linguistic entities; I take them to be respectable entities in their own right... Since I cannot believe that I and all my surroundings are a set of sentences (though I have no argument that they are not), I cannot believe that other worlds are sets of sentences either.

(Lewis 1973: 85–86)

The strong ontological commitments made by modal realism have been far from universally accepted. Kripke (1980) rejects the view that possible worlds are not qualitatively dissimilar the actual world and have an existence

\(^3\)For a good overview of this topic, see Hale (1997).
independent of it; rather they are non-actual ‘states of the system’ or “ways the world might have been” (p. 18). Kripke appears to accept the need for partiality in counterfactual thoughts:

What do we mean when we say ‘In some other possible world I would not have given this lecture today?’ We just imagine the situation where I didn’t decide to give this lecture or decided to give it on some other day. Of course, we don’t imagine everything that is true or false, but only things relevant to my giving the lecture; but in theory, everything needs to be decided to make a total description of the world. We can’t really imagine that except in part; that, then is a possible world.

(Kripke 1980: 44)

The clarity of the final sentence suffers from some anaphoric ambiguity, but Kripke’s position seems to be this: to identify the appropriate state of affairs in which to evaluate a counterfactual, we do not compare other maximal worlds with regard to overall similarity; rather, we identify some entities in this world (e.g. the speaker and his match) and stipulate changes to those entities so that the antecedent is made true (the speaker strikes the match). Evaluating the counterfactual involves considering the repercussions of the changes we have made (whether the match lights or not). In logic, possible worlds must be maximally described; for example, the number of leaves on every tree on every planet must be specified. Lewis is committed to accepting that counterfactuals refer to maximal worlds; Kripke asserts that ‘ministates’ or ‘miniworlds’ are all that is required (p. 18).

Kripke’s internalist position leads to an interesting problem. He places a limit on the changes that may be stipulated of entities by requiring that their ‘essential properties’ may not be altered. The exact nature of essential properties is not stated, but it is clear that he does not allow a counterfactual like (34)5:

(34) If I were a bird, I would be happy.

4To use Kripke’s colourful analogy, possible worlds are not “discovered by powerful telescopes” (p. 20).
5cf. p. 46: “Supposing Nixon is in fact a human being, it would seem that we cannot think of a possible counterfactual situation in which he was, say, an inanimate object; perhaps it is not even possible for him not to have been a human being.”
Being human is an essential property of me, which my being a bird would clearly contravene. Hence it is impossible to stipulate my being a bird. Yet the sentence seems completely acceptable (to me, at least), and it could be used in a fairly normal conversation. Furthermore, it means exactly what it says, how it would be if I were in fact a bird. I do not know how it would come about, but I can talk about how it might be. (34) does not discuss how it would be if I suddenly turned into a bird, or if I acquired some birdlike properties. So Kripke cannot allow (34), although it seems that we wish to accept it.

Might mental models be possible worlds? As worlds are traditionally conceived, it is clear that they cannot be. Possible worlds are maximally specified, and contain far more facts than could ever be known. On the other hand, both mental models and possible worlds represent alternative ways that things could be. Mental model theorists have been ambivalent regarding possible worlds: Johnson-Laird (1983) states that “universes of possible worlds are metaphysical luxuries that have nothing to do with the way in which people ordinarily understand conditionals” (p. 58), but he also recognises that mental models correspond to a fragment that is consistent with many possible worlds. A recurring idea in the mental models literature is that of arbitrary representation. When we consider the premise If there is a circle, there is a triangle, we create a set of mental models that contain an arbitrary triangle, which stands for any possible triangle; when we consider All men are handsome, we create a set of models that contain a number of arbitrary men, which stand for all men. The representation of a state of affairs as a model is not arbitrary in the sense that it is fully specified, but some of its contents are irrelevant; rather, it is underspecified, and can stand for any state of affairs in which the facts of the model hold. Byrne (1996) unambiguously claims that the mental model theory of counterfactuals is “essentially a psychologically constrained version” of possible-worlds analyses (p. 162).

If possible worlds were partialised, such as Kripke might be understood to suggest, we might could arrive at ‘miniworlds’ that are more like mental models. After all, the main argument against possible worlds is their maximality. Barwise (1989: ch. 4) considers a similar suggestion by Cresswell (1988). Cresswell proposed that restricted worlds that settle all issues in only a small subspace of the actual world would be able to do everything that a situation can, whilst retaining possible worlds semantics. Barwise observes that if conventional possible worlds semantics are to be retained,
then the issue space of such a ‘small world’ will be by necessity rectangular, in that every relation in the world must be defined for every object in the world. This rules out the possibility of representing, for example, that some things in the small world are green, and that there is a car in the world, but it is unspecified whether the car is green or not. The partial representation offered by such small worlds therefore does not seem partial enough for our purposes, as this is just the kind of situation that we would like to represent.

Might mental models be situations? Both modes of representation clearly share a concern with partiality. Like situations, mental models only contain some information, which is a subset of all our information or beliefs about the world. Unlike Cresswell’s small worlds proposal, situation semantics does not require a rectangular issue space. Mental models can be nested (Johnson-Laird 1983: ch. 10), as can situations; it is nonsensical to nest one possible world inside another\(^6\). Situation semantics and mental model theory possess similar conceptions of truth (cf. section 4.5). Yet mental model theorists have had little to say about situation semantics (with the exception of a few approving comments in Johnson-Laird 1983), and situation semanticists have had little to say about mental models (besides a few approving comments by Barwise 1993, and disapproving comments by ter Meulen 1993). Recalling Johnson-Laird’s conception of mental models as ‘fragments’ consistent with a large number of possible, it seems that mental models are not so much analogous to situations as to situation types. A model such as \( \triangle \Box \) does not stand for a single state of affairs where there is a triangle and a circle; rather it stands for the set of states of affairs where there is a triangle and a circle. It is obviously wrong to say that mental models are situation types, but it is interesting to consider how the two theories in question might converge.

The situation semantics and mental models accounts of counterfactuals are framed in dissimilar terms; the former posits conditional constraints, and the latter posits sets of models. Once we identify mental models with situation types, the dissimilarity fades. We consider the fully explicit set of models for \( \text{if } p, \text{ then } q \):

\[
\begin{array}{cc}
p & q \\
\neg p & \neg q \\
\neg p & q \\
\end{array}
\]

In this model set, \( p \) only holds in the first model; hence this model represents

\(^6\)Even for small worlds, which are such that “it is possible that they are all that there is”, according to Cresswell.
all states of affairs in which \( p \) holds. Therefore, the model set states that any state of affairs where \( p \) holds is also one where \( q \) holds. In the situation semantic account, the conditional constraint \( P \Rightarrow Q | B \) holds that for any situation of type \( P \) (the class of situations where \( p \) is the case) there is a compatible situation of type \( Q \) (the class of situations where \( q \) is the case), so long as the situation of type \( P \) is also of type \( B \). The reference to compatibility (as opposed to identity) assures an even greater degree of partiality for the situation semantics account, but the analyses are equivalent. We must then ask what component of a mental model correlates to the set of background conditions \( B \). Most importantly, any additional information in the model(s) where the antecedent and consequent of the conditional both hold must be part of the background conditions. It is unclear what is conveyed beyond the contents of these models, apart from a vague *ceteris paribus* condition. For example, the model set in 1(a) might be equivalent to the conditional constraint in 1(b), and the model set 2(a) might be equivalent to the conditional constraint in 2(b):

1. (a)        \[ \text{struck(match)} \quad \text{lights(match)} \]
   \[ \neg \text{struck(match)} \quad \neg \text{lights(match)} \]
   \[ \neg \text{struck(match)} \quad \text{lights(match)} \]

(b) \( S \Rightarrow L | B \)
   \[ S = [s| \text{in } s : \text{at } l : \text{struck, match};1] \]
   \[ L = [s| \text{in } s : \text{at } l : \text{lights, match};1] \]
   \[ B = [s| \text{normal, } s;1] \]

2. (a)        \[ \neg \text{oxygen} \quad \text{dry(match)} \quad \text{struck(match)} \quad \text{lights(match)} \]
   \[ \neg \text{oxygen} \quad \text{dry(match)} \quad \text{struck(match)} \quad \neg \text{lights(match)} \]
   \[ \text{oxygen} \quad \neg \text{dry(match)} \quad \text{struck(match)} \quad \neg \text{lights(match)} \]

   \[ \ldots \]

(b) \( S \Rightarrow L | B \)
   \[ S = [s| \text{in } s : \text{at } l : \text{struck, match};1] \]
   \[ L = [s| \text{in } s : \text{at } l : \text{lights, match};1] \]
   \[ B = [s| \text{in } s : \text{oxygen} \land \text{dry, match} \land \text{normal, } s;1] \]

Both 1 and 2 above license the conditional *If I strike this match, it lights*. A similar analysis (with appropriate footnotes on the models) would describe the corresponding counterfactual. The definition of a ‘normal’ situation will
depend on factors that are not available in the speaker’s set of models (although they will be manifest in his general knowledge and in the actual situation), and are certainly not available to the hearer. The hearer will most likely interpret ‘normal’ as corresponding to his own set of beliefs, unless those beliefs do not support the conditional constraint; in this case, he will have to play the conditional guessing game to establish what the speaker has in mind.

We cannot ask whether possible worlds or situations have counterparts in mental logic, but it is reasonable to ask what kind of logic mental logic might be. The theory presented in Braine & O’Brien (1998a) has not been fully developed beyond the domain of propositional reasoning, but the ‘true’ system of mental logic (should it exist) must be far more expressive. We would hope that mental logic is in principle sound, disregarding the problems of nonmonotonicity. The limitations on the use of recursive derivation rules and the constraints on the Schema for Conditional Proof seem to suggest that it is not complete – a conclusion may follow from a set of premises, but fail to be derivable.

4.4 On Conversation

All naturally occurring counterfactuals are anchored in context. Frequently, the context is contained in a conversation between two or more parties. In section 2.5.2 we considered how the cooperative nature of well-run conversation requires that the speaker make as explicit as possible (within reason) the conditions that he believes license the counterfactual, and that the hearer do his best (within reason) to identify conditions that license the speaker’s assertion. Conversations are rarely, if ever, ideally run, and counterfactuals are more likely than other assertions to highlight suboptimal cooperation. When the grounds for a counterfactual assertion are particularly unclear, e.g. for an assertion *If this table could fly, it would fly to the Moon*, the hearer is left to guess what the speaker has in mind. If no rational grounds are identified, the hearer must deny or question the counterfactual – though to deny it, he should have some grounds for saying what the table would or would not do if it could fly – or else simply accept its truth.

---

7 Although a tentative outline of ‘mental predicate logic’ is given by Braine 1998.
8 The fact that people frequently disagree about the truth of counterfactuals seems to be conclusive evidence that counterfactuals are not usually evaluated with regard to some
We also saw that failures of certain inference rules can hinge on covert shifting of the ‘language score’. In mental model theory, these rules become invalid when the conclusion does not follow from the integrated model of the premises. Of course, this in turn depends on the interpretation of the premises, of which the theory does not provide an account. An advance in this area might be suggested by the congruences with situation semantics that were identified in the previous section. If an asserted counterfactual contradicts the perceptual or general knowledge most available (in Tversky & Kahneman’s sense) to the hearer, then if he is not to deny the counterfactual outright, he must identify the conditions that might license it, and represent them explicitly in his models of the assertion; otherwise no background conditions need to be explicitly represented.

4.5 On Truth

It may be true in a metaphysical sense that if the table I am sitting at could fly, it would fly off to the Moon and take my computer with it; alternatively, it may not be true. Either way, its absolute truth, falsity or indeterminacy is unknowable. We cannot line up all the relevant possible worlds and compare their similarity. The truth of a counterfactual (or any other entertainable proposition) is, however, judgable, in that we can decide whether or not there are grounds for placing our belief in its truth. Formal semanticists have generally followed Frege in discussing truth as an abstract notion, but a number of recent theories have proposed more psychologically plausible conceptions of truth. Jackendoff (1996) describes a theory of I-Semantics, which is concerned with the cognitive nature of meaning, just as Chomsky’s theory of competence relates to internalised knowledge of syntactic structure. He offers the following definition of truth in I-Semantics (p. 543):

A speaker of language $L$ judges sentence $S$ true iff conditions $C_1, \ldots, C_n$ obtain in his or her construal of the world (subject to limitations of memory and processing, and under an idealization of uniformity among speakers.

The term ‘construal of the world’ is certainly compatible with ‘mental model of the world’. As was observed in the discussion of Byrne & Tasso’s canonical or ‘standard’ notion of similarity.
(1999) experiments, models do not represent truth, merely ‘possibilities’. This is clearly desirable; of the three models for if $p$, then $q$, no more than one can represent the true state of affairs. A mental model is true if and only if it corresponds to the way things are in the actual world. To assert that if $p$, then $q$ is true is to assert that one model of its interpretation corresponds to the actual world – that the entities in the actual world that correspond to the tokens in the model have the same properties as the corresponding tokens do, and stand in the same relation to each other as the corresponding tokens do in the model (Johnson-Laird 1983: ch. 10). But we cannot verify or falsify an assertion by checking the actual external world, as we have no direct access to its facts. Again, we must compare its interpretation to our ‘construal of the world’; this may be the stock of available beliefs that manifests itself in a reasoning-type model, or it may be our internal perceptual model of the world\textsuperscript{9}. A consequence of this concept of truth is that it seems to implement a ‘closed world assumption’; if a possibility does not appear in a fully explicit set of models, then it is judged to be false.

As we have already observed, a mental model stands for very many ways the world could be. For example, any model is compatible with an infinitude of traditional Tarskian models. Searching a model set for counterexamples is a cognitively constrained version of proving invalidity via the construction of a countermodel (cf. Gamut 1991: sec. 4.2). The correlation between mental models and situation types is also reflected in the conceptions of truth they support. According to the situation semantics account, to assert that an utterance (or discourse) is true is to assert that the situation being referred to is of the type described by the utterance. According to mental models theory, to assert that an utterance (or discourse) is true is to assert that one of the models in its interpretation describes the situation being referred to. The only significant difference is that shown up by the phrases the type described by the utterance and one of the models in its interpretation, and that is no difference at all. The assertion if $p$, then $q$ corresponds to a single situation type, but to a set of three models; however, the interpretations are functionally equivalent.

The use of arbitrary representatives in mental models suggests a parallel with the use of discourse referents in Discourse Representation Theory (usually abbreviated as DRT; described in Kamp & Reyle 1993, van Eijck &

\textsuperscript{9}Marr’s (1982) inspirational theory of vision describes how perceptual information is used to construct what is essentially a mental model of the world around us.
Kamp 1997). DRT uses an intermediate form of representation to capture the dynamic nature of discourse – as a discourse progresses, a Discourse Representation Structure (DRS) is incrementally constructed. Indefinite noun phrases such as a man introduce new discourse referents into the domain of discourse, and the DRS for the sentence A man walks is true in a model iff the discourse referent introduced by the NP a man can be assigned to an entity in the model such that that entity walks. Similarly, the sentence There is a triangle introduces an arbitrary representative triangle $\triangle$ into a mental model. The model is true with regard to a reasoner’s ‘construal of the world’ if that representative triangle can be matched to some specific triangle in that construal of the world. Johnson-Laird (1983: 439) acknowledges his debt to Kamp’s (1981) early presentation of DRT in this regard.

What ‘construal of the world’ is appropriate for judging the truth of counterfactuals? We recall Johnson-Laird & Byrne’s (2002: 652) remark that “One cannot observe counterfactual states, and so the truth or falsity of a counterfactual conditional about a contingent matter may never be ascertained.” This does not explain how we consider some contingent counterfactuals to be true and some to be false. It seems reasonable to suggest that we evaluate a given counterfactual with regard to the incomplete set of mental models that its interpretation yields, as modulated by our most available beliefs on the matters it refers to. This reflects the approach that I have advocated in this chapter: to evaluate If I had struck this match, it would have lit, we construct the set of mental models yielded by our interpretation of the counterfactual, and we integrate into the set those salient beliefs that by some mechanism ‘leap out’ at us. In the previous chapter, we considered how certain beliefs may become more available than others; in any match-striking scenario, the belief that struck matches generally light is highly salient, but the salience of the belief that the match must be dry if it is to light will depend on factors such as prior experience or some perceptual or conversational clue that indicates a possible lack of oxygen. If the consequent still holds in the integrated model, then we assent to the counterfactual. If it does not, then we might try to revise our beliefs in an acceptable way so as to verify the conditional, or else we might simply deny its truth. My proposal is in fact rather different from the traditional Ramsey Test analysis, in that both antecedent and consequent are represented in the mental model before the final evaluation is made. We do not simply add the antecedent and see what follows. This has two pleasing side-effects. Firstly, the analysis explains how abstract conditionals that do not clash with any
of our prior beliefs are generally accepted as true if they are asserted – the lack of any verifying beliefs may lead us to question the speakers’ grounds for the assertion, but we should not deny it. Secondly, it offers an explanation of the ‘biased similarity criterion’ I proposed in section 2.5.2 to account for counterfactuals like *If I were taller, I would be good at basketball.* It is not necessary to check the truth of the consequent in all sufficiently similar states of affairs; it is sufficient that none of our salient beliefs contradict it. A similar account can explain Chisholm’s Apollo sentences. To paraphrase Johnson-Laird & Byrne (1993b): Humans are falsificationist in theory, but verificationist in practice.

### 4.6 Conclusion

“With no effort, he had learned English, French, Portuguese and Latin. I suspect, however, that he was not very capable of thought. To think is to forget differences, generalise, make abstractions. In the teeming world of Funes, there were only details, almost immediate in their presence.”

– Jorge Luis Borges, *Funes the Memorious*

In Borges’ story, the memorious Funes could not apprehend the underlying structure of the decimal number system, and created a superior system of his own. He assigned each number a unique name that bore no relation to that of any other number. For example, *Máximo Pérez* meant 7013, and *The Railroad* meant 7014. He extended his system to describe more than twenty-four thousand numbers, and remembered the name of each.

The experience of Funes was almost the opposite of the human experience. Our cognitive capacity is generally very limited, and we do not assign every unique entity a unique name. We do not have a distinct word to refer to each distinct table; instead, we use the general term *table*, which occludes the differences among its referents. Our number system, which so baffled Funes, contains just 10 symbols and some simple combinatory rules. The ability to abstract and generalise is central to our experience of the world, and ultimately to our cognitive processes. Counterfactual reasoning lies near the apex of our abstractive talents. When we think counterfactually, we do not just identify relations in the world around us; we abstract from those relations to identify relations in situations that do not exist. Generalisation has a price; gains in efficiency are balanced by losses in specificity. The price
is usually an acceptable one, yet over-vague counterfactuals can sometimes test the limits of this acceptability.

What are we then to conclude about counterfactuals? It would surely have been a waste of 122 pages if all we had established were that counterfactuals are vague and difficult to analyse. I believe that this is not the case. I hope to have demonstrated that in spite of their differing aims and methodologies, philosophy and psychology have faced similar issues in their investigations of counterfactuals. I have defended a number of claims about the way we use counterfactuals and how they should be analysed; the most important are the following:

1. I have defended the thesis that we generally judge counterfactuals to be true or false, given a minimum amount of ‘relevant information’, though we cannot know whether they are ‘really’ true or false from a metaphysical standpoint. If a counterfactual is overly vague, we may not be able to reach a verdict, but this will only be the case when insufficient contextual information is available. This is similar to the case of vague predications such as \( x \text{ is flat} \), which without any information about a criterion of flatness seem indeterminate or even false. Some may doubt that any contingent counterfactual can be assigned a truth value. I cannot prove such doubters wrong, but nor can they prove me wrong. Intuition must decide.

2. Most interesting problems regarding counterfactuals resolve to the problem of identifying the information that licenses judgements of their truth or falsity. In philosophical analyses, this was manifested as the problem of similarity in possible worlds accounts and the problem of relevant conditions in metalinguistic accounts. In psychological analyses, the vague notion of ‘availability’ was seen to be crucial. Counterfactual truth judgements depend on a reasoner’s prior ‘stock of beliefs’, and which beliefs he is willing to abandon in order to consider a non-actual scenario.

3. Any cognitively plausible theory of counterfactuals must incorporate the principle of partiality. When we consider a counterfactual, we cannot consider all the relevant facts of the world that could affect its truth or falsity. We frequently fail to recognise the consequences of our beliefs, and appear to reason in an irrational manner. The concept of bounded rationality is necessary to express the idea that we are as
rational as we can possibly be, given our basic cognitive limitations. Humans are not ideal reasoners.

4. The rooting of counterfactuals in a discourse context is essential for the resolution of their inherent vagueness. Problems can arise when one party in a conversation disregards the principle of cooperation and substitutes one criterion of vagueness for another without informing the other party.
Chapter 5

Epilogue: Computational Perspectives
5.1 Counterfactuals in Artificial Intelligence

Researchers in artificial intelligence have paid attention to counterfactuals because of their relationship to McCarthy & Hayes’s (1969) famous Frame Problem. The Frame Problem is essentially the problem of describing all aspects of the world that are left unchanged by a given action. Given a suitably small domain, it is possible to exhaustively specify all the non-effects of each possible change in the world by means of frame axioms. However, this approach is untenable when the domain is scaled up, as the number of requisite frame axioms is subject to combinatorial explosion. Intuitively, it seems that we need to appeal to some common sense law of inertia, which states that “Nothing changes, except when there is explicit evidence to the contrary” (Hamm & van Lambalgen 2000). On a general level, such a law is analogous to a method of determining the relative similarity of possible worlds, as both involve the identification of those aspects of the world that are affected by a particular change, and of those aspects of the world to which the change is irrelevant.

Counterfactuals in themselves have not been applied in computer science to a large degree, presumably due to the formal difficulties they pose. One notable exception is Costello & McCarthy (1999). A Lewisian account of causation has been used by O’Neill (2002) for identifying causal relations in databases.

Fritz Hamm and Michiel van Lambalgen have advocated the use of formalisms from artificial intelligence in the analysis of natural language semantics (Hamm & van Lambalgen 2000, van Lambalgen & Hamm 2001). In Hamm & van Lambalgen (2000) they apply techniques initially developed to tackle the Frame Problem to accounts of nominalisation and the progressive aspect in English. In view of the affinity between the Frame Problem and the problem of counterfactual validity, it would be interesting to consider whether such techniques might successfully formalise the semantics of counterfactuals.
Bibliography


Boole, George. 1854. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. London: Macmillan.


BIBLIOGRAPHY


n´ı Bhriain, Áine. 2002. Propositional reasoning via the construction of mental models. BA (Mod.) Dissertation, Department of Computer Science, Trinity College Dublin.


BIBLIOGRAPHY


van Lambalgen, Michiel & Fritz Hamm. 2001. Moschovakis’ notion of meaning as applied to linguistics. To appear in M. Baaz, J. Krajicek (eds.), Proceedings of Logic Colloquium ’01; available online at http://www.semanticsarchive.net/Archive/GFmYWFjZ.


