Hope and Understanding: Aspects of a Theory of Rational Inquiry

Thesis submitted for the degree of Doctor of Philosophy in the University of Sheffield

Nigel Gibbions, Department of Philosophy, June 2000

What we conquer are the small things,
And victory itself makes us small.
The eternal and uncommon
Does not want to be shaped by us.
This is the Angel who to the wrestlers
Of the Old Testament appeared:
When the sinews of his adversaries
In the battle stretch metalically
He feels them under his fingers
Like the strings of profound melodies.

Who was overcome by this Angel,
Who has often abstained from battle,
He walks justified, upright,
And proud of that hard hand,
Which, as if moulding, gently encloses him.
Victories are not inviting to him.
His gain is to be profoundly vanquished
By ever greater things.

[Rainer Maria Rilke]

Hope and Understanding: Aspects of a Theory of Rational Inquiry

Summary of Thesis

The thesis aims to establish two claims that have a role to play in any theory of rational inquiry:

- The aim of rational inquiry is understanding; this is achieved to the extent that we end up with an adequate model of the object of our inquiry.
- Rational inquiry presupposes fundamental hopes in the orderliness of the universe, and in our ability to discern that order, which it is legitimate for us to entertain.

Chapter 1 sets the scene by describing how the two main themes of the thesis - hope and understanding - are treated in Plato's *Meno* dialogue, and in Peirce's more diffuse writings on the nature of scientific inquiry.

Chapter 2 is a detailed analysis of the nature of hope and its role in our inquiries. The chapter aims to answer two questions - what is hope; and when is hope legitimate - and to apply those answers to the specific hopes that underlie the possibility of rational inquiry.

Chapters 3, 4 and 5 are concerned with understanding, the goal of inquiry. Chapter 3 makes some important initial distinctions, and presents an analysis of *dynamic* understanding (which occurs at a particular moment in time). The chapter also introduces the concept of a model, and shows how it may be used to shed light on the nature of *static* understanding (which persists over a period of time).

Chapter 4 presents an argument from analogy: understanding and spatial orientation have many features in common; the best way of accounting for this is to assume that understanding a domain involves having a map-like representation - a model, in other words - of it.

Finally, chapter 5 argues that current work on the nature of scientific theories and scientific explanation provides further evidence for the claim that understanding a given domain is a matter of having a model of it.

Nigel Gibbions, Department of Philosophy, June 2000

Acknowledgements

I would like to thank all the people who have supported me, during the writing of this thesis.

Firstly, my supervisor, Chris Hookway, for his patience, and willingness to give my ideas a fair hearing, even when they were at their most inchoate. Without his advice, comment and criticism, this would be a much poorer piece of work.

Steven Makin supervised me during the first year of this Ph.D, when it was still supposed to be about Wittgenstein's discussion of rule following. I thank him for his help, and for encouraging me to explore the interest in the Meno paradox that eventually hijacked the thesis.

Chris Bennett and Christopher Goodman deserve special thanks for introducing me to relevant and stimulating papers that I might otherwise not have read at the*Phronesis* reading group.

I would also like to thank the staff and graduates in the Philosophy Department at Sheffield, for working together so effectively, to create an excellent environment in which to do a Ph.D: challenging and supportive in just the right measures.

On a personal note, I would like to thank my family and friends for their support throughout this project (and before and after it, for that matter), with a special mention to Tim, Kristin, Betty-Ann and Keith. I would also like to thank the people who gave me not just a roof over my head, but a home, while I was writing this thesis: Sue, when I first came to Sheffield, and, later on, Steph, Danny and Lindiwe.

Table of Contents

List of Figures and Tables

Introduction

This is a thesis about the nature of rational inquiry. But, given the size and complexity of the topic, it would be unrealistic to aim to present a comprehensive philosophical theory of rational inquiry, in a work of this scale. Instead, I aim to establish two claims that will play an important role in any such account:

- The aim of rational inquiry is understanding; this is achieved to the extent that we end up with an adequate model of the object of our inquiry.
- The possibility of rational inquiry presupposes fundamental hopes, which, in the circumstances, it is legitimate for us to entertain.

The meaning of these claims will become clearer as the thesis develops.

Chapter 1 introduces the two main themes of the thesis, hope and understanding, and describes the way these themes are dealt with by two major philosophers: Plato, in the*Meno* dialogue, and Peirce, in a number of more diffuse writings on the nature of scientific inquiry. From these two sources, which, despite the two millennia gap between them, tell remarkably similar stories, it is possible to glean some tantalising hints about the aim of inquiry, and the role that hope might play in a theory of rational inquiry.

Chapter 2 begins the work of constructing a detailed defence of the two key claims of the thesis, by developing the ideas about the role of hope in rational inquiry tentatively proposed in chapter 1. It focuses upon two questions:

- What is hope?
- When is hope legitimate?

In each case, chapter 2 suggests both general answers, and answers that are relevant to the specific, and rather unusual, hopes presupposed by the possibility of rational inquiry. In its search for those answers, it draws upon the work of Aquinas, William James, and Kant. The chapter reaches the reassuring conclusion that we *may* legitimately entertain the hopes that underlie the possibility of rational inquiry.

Having explored the issue of hope and its legitimacy at length, in chapter 2, we turn to the second theme of the thesis, understanding, in chapters 3, 4 and 5. Since this is a rather amorphous concept, chapter 3 begins by introducing a number of important distinctions, clarifying the object of our inquiry in the process. Much of the remainder of the chapter is then taken up with an analysis of the nature of what I call *dynamic* understanding, which, roughly speaking, is the *episodic* aspect of understanding. Towards the end of the chapter, however, I turn to what I call *static* understanding, which, roughly speaking, is a state of the individual that persists over time[1]. At this stage, I introduce the concept of a model, and show how certain features of the *Meno* dialogue suggest that we have static understanding of a given domain to the extent that we have an adequate model of it. The purpose of chapters 4 and 5 is to provide further evidence for this, the second key claim of the thesis.

Chapter 4 begins by noting the prevalence of spatial metaphors in the language we use to describe our inquiries, and suggests that this is a clue to the nature of understanding[2]. In following up this clue, I develop an analogy between spatial orientation and understanding, and argue that the former is a matter of having a map-like representation of one's environment, a map being a model of the terrain it represents. Taken together, these conclusions form an argument from analogy: understanding and spatial orientation share many features; and the best way of accounting for these similarities is to suppose that understanding a given domain involves having a map-like representation - a model, in other words - of it.

Chapter 5 argues that further evidence for the view that understanding a given domain is a matter of having an adequate model of it, is provided by current work on the nature of scientific theories and scientific explanation. Scientific theories are prime sources of scientific understanding, and, according to one influential philosophical account, may also be thought of as providing us with models of the real physical systems that fall within their intended scope. Scientific explanations are also sources of scientific understanding, and, in the second half of chapter 5, I show how the concept of a model may be used to make sense of the current debate about the nature of scientific explanation, suggesting that the three main accounts may not be rivals after all.

[1] See pp80-83 for a fuller explanation of the distinction between dynamic and static understanding.

[2] For the rest of this introduction, the term "understanding" (without qualification) should be taken to refer to static understanding.

As the reader will have gathered from this brief summary, the scope of the thesis is wide. For this reason, I should make it clear at the outset that the thesis is not intended as an in depth investigation of Plato's theory of inquiry in the *Meno*, or of Kant's writings on the nature and legitimacy of hope, or of the current debate about the nature of scientific explanation. It is fair to say that no single topic covered in the thesis receives the individual attention it deserves, in the sense that any one of them could be the topic of an entire thesis in its own right. The interest, if any, of the thesis is in the way it pulls all of these topics together, and uses them to build up a coherent image of a fragment of a philosophical theory of rational inquiry. This approach may engender frustration when topics have to be skipped over quickly - the section on Kant's work on the "needs of reason"[3] probably suffers the most in this respect - but the wider perspective it offers is, I believe, sufficient compensation.

[3] See section 2.4.

Chapter 1: Plato and Peirce on Rational Inquiry

1.1 Introduction

The purpose of this chapter is to introduce the two main themes of the thesis, hope and understanding, and to begin to describe their place in an account of rational inquiry. I do this by exploring the treatment of these themes in two texts: Plato's *Meno*, and Peirce's more diffuse writings on scientific inquiry. Although these works are separated by a couple of millennia, I hope to persuade the reader that there are unexpected and interesting parallels between their accounts of inquiry.

I make a start on this task in section 1.2, where I present an overview of the *Meno* dialogue, zooming in at especially interesting points. Any discussion of the dialogue must say something about the paradox it features, which purports to undermine the possibility of rational inquiry. I agree with most commentators, in believing that the paradox is rather easily disarmed. However, I think that there are a number of problems in its immediate vicinity that, although not so clearly articulated as the paradox, are rather harder to resolve. Peirce's work on inquiry, which I examine in the section 1.3, confronts these issues directly, and openly suggests solutions that are remarkably similar to those only hinted at in the *Meno*.

1.2 Inquiry in the *Meno*

In this section, I present an overview of Plato's *Meno*, the first work of Western philosophy, as far as I know, to raise interesting questions about the nature of inquiry. In doing so, I aim to begin to establish three claims that will be explored in greater depth in chapters 2 to 5:

1. The dialogue makes more sense if the Greek word *episteme*, which describes the goal of inquiry, and which is usually translated as *knowledge*, is translated as *understanding*.

2. Understanding, the goal of inquiry, requires a distinctive intellectual effort on the part of the inquirer; the reward for this effort is a kind of *stability* in one's beliefs.

3. In order to make the commitment to serious inquiry described above, it is necessary to entertain hopes in things that are not amenable to rational demonstration.

The structure of the dialogue is quite complex, but, for our purposes, the following rough analysis will suffice:

1.	The acquisition of virtue (I).	70a-70c
2.	The priority of definition.	71a-71b
3.	Attempts to define virtue.	71c-79e
4.	Meno's paradox.	80a-81a
5.	The Theory of Recollection.	81b-81e
6.	The slave boy dialogue.	82a-86c
7.	The acquisition of virtue (II).	86d-96d
8.	True opinion vs knowledge.	96e-100c

I shall look at all of the above episodes, although depth of coverage will be uneven, reflecting the main concerns of the thesis. In particular, I have little to say about item 7, a long stretch of the dialogue during which it appears that Socrates gives in to Meno's demand that his original question about the origins of virtue be answered.

1.2.1 The Acquisition of Virtue (I)

Many of Plato's dialogues have lengthy preambles, some incorporating elaborate framing devices. In the *Symposium*, for example, Apollodorus recalls telling the tale - as told to *him* by Aristodemus - of a famous symposium at which Socrates was present to Glaucon at a recent meeting. The symposium itself took place many years before the meeting, when both Apollodorus and Glaucon were children. In contrast, the *Meno* begins extremely abruptly: upon finding Socrates in the square, Meno, a young aristocrat from Thessaly, visiting Athens, immediately asks:

> Can you tell me Socrates - is virtue something that can be taught? Or does it come by practice? Or is it neither teaching nor practice that gives it to a man but natural aptitude or something else[1]?

The abruptness of Meno's opening question is significant, for it indicates something about his conception of education, a conception that imposes severe intellectual limitations upon him. This theme is developed in the remainder of section 1.2, and in chapter 3[2].

[1] Plato, *Meno*, 70a.

[2] See especially chapter 3, pp106-108 and p116, for a diagnosis of Meno's intellectual shortcomings.

1.2.1 The Priority of Definition

Socrates does not answer Meno's question about the origins of virtue directly. In fact, Socrates says that he *cannot* answer the question, for he does not know what virtue is; and, he claims, until one knows what something *is*, it is futile to inquire further into its nature. To illustrate his point, he asks, rhetorically, if it is possible to say whether Meno is handsome or rich, without first knowing who Meno is[3]? Socrates' insistence, in this and other dialogues, upon the priority of definition has puzzled many readers of Plato. In this section I argue that it is possible to assign a plausible interpretation to Socrates' claim. However, in order to do so, we must translate *episteme* as *understanding*. Exploring this issue will also reveal connections between Socrates' conception of inquiry and Peirce's. Let us begin by distinguishing between three kinds of definition: linguistic, conceptual and real[4].

A *linguistic* definition supplies information about how a given term, T, is used, one that could be used to introduce the word to someone who was unfamiliar with it. Typically, this is accomplished by citing a term synonymous with T that is more familiar to the inquirer. A reasonable linguistic definition of "celerity", for example, would be "speed". Most dictionary definitions are linguistic in this sense. Such definitions are relatively easy to formulate and grasp, but also tend to be somewhat superficial.

Conceptual definitions dig deeper, in assuming that the inquirer already knows how to use the term in question, but is trying to relate the concept it expresses to other concepts she has grasped. For example, speed could be defined in this manner as distance travelled, divided by time taken[5]. Analytical philosophy is a search for conceptual definitions, insofar as it is concerned with providing lists of necessary and sufficient conditions for philosophically important terms[6].

[3] *Ibid.* 71b.

[4] This classification and the following discussion is based on I. M. Crombie's, "Socratic Definition". It may be thought that there is another type of definition that is especially relevant to the virtue case; namely one that describes the *rules* a competent user is following (however unreflectively) in applying a term. However, Socrates regards *all* definitions (linguistic, conceptual or real) as saying something about how the use of a term is regulated - as is shown by his insistence, throughout the dialogue, that candidate definitions of virtue and the like are checked against one's prior intuitions about the use of the term. A description of the rules governing the use of a term is not an additional type of definition; rather it is something that is already implied by other types of definition.

[5] Strictly speaking this is a definition of *average* speed over the given interval.

[6] Of course, this is very crude; it is *not* intended as a definitive description of the nature of analytical philosophy.

Finally, a *real* definition is an attempt to describe the essence of something. This kind of definition has much in common with the efforts of natural scientists to say what, for example, electricity really is. Real definitions may be of two types. In the first, a class of entities is identified as a potential natural kind, and the inquirer attempts to identify some common feature which explains our natural tendency to group these items together. In the second, the inquirer already has in mind some property or feature - "the cause of cancer", say - and seeks to determine what in the world satisfies this description. The former type of real definition moves from extension to intension, whilst the latter proceeds in the opposite direction.

The question now arises: what kind of definition is Socrates after in the*Meno*? The best way to answer this is to see what the dialogue rules out. Firstly, it is unlikely that Socrates is asking Meno for a linguistic definition, since both participants in the dialogue use the term "virtue" with assurance. Also, it is difficult to see how a purely linguistic definition of virtue could be of any philosophical interest. Secondly, if it is a real definition of virtue that Socrates is after, certain exchanges suggest that it is of the first type, from extension to intension, rather than the other way around. For example, Meno's first attempt to define virtue is just a list of virtues, each appropriate to different groups of people (men, women, children, old people, slaves, and so on)[7]. Socrates' response is to insist that Meno identify the respect in which all of these cases "don't differ at all, but are the same"[8]. Here, and throughout the dialogue, it is taken for granted that Meno is capable of identifying*examples* of virtue, and that Socrates concurs with his judgements. The problem is to say what property or properties all of these particular cases share[9].

This leaves us with two possibilities: either Socrates is after a real definition of virtue that takes us from the extension of the term to its intension, or he seeks a conceptual definition. My feeling is that Plato does not distinguish clearly enough between these two options in the

[7] *Ibid.* 71e-72a. For more on Meno's attempts to define virtue, see section 1.2.3 below.

[8] *Ibid.* 72c.

[9] This is an appropriate point to insert a brief note on the relevance of "family resemblance" considerations to questions of definition. Do Wittgenstein's remarks (*Philosophical Investigation*, section 67 and elsewhere) on the topic mean that Socrates' demand cannot be met? I think that this depends upon the topology of the concept in question. One possibility is that there is a hard core to the extension, with a defining set of essential properties; another is that, although there is no such hard core, it is possible to identify a small number of "axes of variation" along which family members may be arranged. Either of these possibilities, is consistent with the spirit, if not the letter, of Socrates' demand. Only if the topology of the concept to be defined is horribly convoluted would the search for a definition be thwarted; but perhaps there are good reasons why few, if any, of our concepts are like that.

Meno. Because of this, I suggest that we try a different tack: instead of scanning the text of the dialogue for evidence that directly supports one reading or the other, let's think about the claim that we must know what something is before we can know anything else about it. Does whatever plausibility this claim has depend upon taking Socrates' demand for a definition in one of our remaining senses rather than the other? Consider the following statements, derived from the principle of the priority of definition by substituting the two remaining candidate senses of *definition*:

1. If one cannot define X in terms of other concepts with which one is familiar, one can't know anything about X.

2. If one doesn't know what the essence of X is, one can't know anything about X's accidental properties.

The second claim sounds fairly reasonable, but can't be quite right. After all, someone could know that gold is yellow and shiny, without knowing what its essential properties - that it has atomic number 79 and so forth - are[10]. But suppose that we translate *episteme* as *understanding*, rather than *knowledge*. The claim then becomes:

2*. If one hasn't grasped the essence of X, one can't understand X's non-essential properties.

This, at least, is a plausible claim, although obviously not an uncontroversial one[11,12]. Unless we have grasped that a hydrogen atom is essentially just an electron orbiting a proton, we are not well placed to understand its more easily observed properties. Similarly, appreciating the true nature of virtue ought to enable us to understand why it is often regarded as admirable or beneficial.

[10] The gold example should bring to mind Locke's distinction between real and nominal essences. See *An Essay Concerning Human Understanding*, III iii, III vi, III x and IV vi 4-9. In "Inquiry in the *Meno*", Fine suggests that having a definition is, for Socrates, like knowing a Lockean real essence.

[11] It may, for example, be an exaggeration to claim that if one hasn't grasped the essence of X, one can't understand *any* of X's non-essential properties. Nevertheless, it's reasonable to suppose that there will be many non-accidental properties of X that one will not be able to account for, unless one has grasped X's essence.

[12] The plausibility of this interpretation is further supported by the fact that Aristotle endorses something like it in *Posterior Analytics* I, 1-10, a work which addresses what we would now call the nature of scientific explanation and understanding. Fine point this out in footnote 13 of her article. For the record, I think that Fine's interpretation of the principle of the priority of definition is essentially correct; she just fails to take the final step I am advocating of reading *episteme* as "understanding".

How does the suggestion that *episteme* be translated as *understanding* play in the alternative version of the priority of definition claim? Substituting into 3, we obtain:

3*. If one does not have an analysis of one's concept of X, one can't know/understand anything about X.

For many values of X, including *hydrogen* and *virtue*, this just looks, at best, like a misleading way of expressing the previous version of this principle[13], and, at worst, simply false. However, for some values of X, this latest version of the priority of definition claim seems quite reasonable. For example, it does not seem so wrongheaded to insist that, unless someone realises that a bachelor is an unmarried man, they can't really understand anything about bachelors. Conversely, knowing the definition of a bachelor enables us to understand why, for instance, so many bachelors live alone.

I conclude that we do not have to make a choice between conceptual and real definitions when formulating the principle of the priority of definition since: (i) Plato did not clearly distinguish between them; and (ii) the kind of definition that is appropriate depends upon the subject matter of the inquiry. But more importantly, as far as this piece of work is concerned, I suggest that we have good reason to believe that the goal of inquiry, according to Socrates, is understanding. Inquiry is, roughly, the search for the essential features of reality, some of which may be hidden, that account for its accidental aspects, many of which are observable. We shall see in section 1.3 that Peirce has a very similar conception of the nature of inquiry.

1.2.3 Attempts to Define Virtue

Meno has three attempts at answering Socrates' question about the nature of virtue before giving up. The first of these, mentioned in the previous section, is just a list of different kinds of virtue. We have already concluded that this is not what Socrates wants; he is looking for the essence of virtue, something that the things that Meno describes have in common, which makes them all examples of virtue. Having taken this point on board, Meno makes two further suggestions: (i) virtue is the capacity to govern men[14]; and (ii) virtue is the desire for fine things, and the ability to acquire them[15]. Both of these efforts are, of course, quite feeble, as Socrates has no trouble in demonstrating. The interesting thing, from our

[13] Describing the atomic structure of hydrogen does not amount to an analysis of our concept of hydrogen.

[14] *Meno*, 73d.

[15] *Meno* 77b.

point of view, is how he does this[16]. In addition, this section of the dialogue ends on a note that is reminiscent of Peirce's account of inquiry.

Socrates challenges Meno: does he think it is virtuous for a slave to govern his master? Meno is forced to concede that it is not, thereby repudiating the proposed definition. This is an example of a dialectical strategy which Socrates adopts throughout the*Meno*, and in other dialogues. The general idea is to show that the views professed by his opponents have implications that conflict with other beliefs that they are unwilling to relinquish. In this case, Meno believes that virtue is the ability to govern men; this implies that it would be virtuous for a slave to govern its master; but this directly contradicts a prior belief that it would not be virtuous of a slave to govern his master[17]. Note that Meno could have arrived at this conclusion himself. If only he had recalled his prior belief that it is not virtuous for slaves to govern masters, he might have realised that his definition of virtue was not consistent with it, and tried to think of something better. In chapter 3, I develop the idea that the ability to appraise information in the light of one's other beliefs is one of the hallmarks of understanding[18]; it is becoming increasingly clear that Meno lacks this ability.

After three unsuccessful attempts to define virtue, Meno concedes that Socrates has reduced him to speechlessness, even though he has spoken confidently about virtue on hundreds of previous occasions to large audiences[19]. He experiences an acute intellectual numbness which he compares to the paralysing effect of the torpedo fish on its prey. This state of intellectual paralysis is called*aporia*. The Greek word *euporia* is literally translated as "easy passage or travel". *Aporia* is the opposite of this; it signifies some difficulty or impossibility of making progress, a roadblock on the path of inquiry, as Peirce might call it[20]. Encountering such obstacles is an almost inevitable result of subjecting one's beliefs to the fiercest possible scrutiny, and is often an acutely demoralising experience. But*aporia* also has a couple of positive aspects: firstly, it teaches us humility in our search for understanding; and secondly, a true thirst for knowledge may arise from the ashes of our conceit. *Aporia* is therefore a transitional, if uncomfortable, stage of genuine, robust inquiry.

[16] For brevity's sake I deal only with the first definition in what follows; similar remarks apply to the other.

[17] The sense in which Meno believes these things is discussed briefly in chapter 3; see p88.

[18] See section 3.4.

[19] *Ibid.* 80a-80b. This is a telling remark. The fact that Meno's beliefs about virtue have never been challenged, even though Socrates has just shown that they are untenable, implies that Meno is far from alone in holding them. Hence, Socrates' invitation to inquire into the true nature of virtue is threatening, for it calls upon Meno to reject the "common sense" views of his fellow citizens. This partly accounts for Meno's reluctance to embark upon a joint inquiry with Socrates.

1.2.4 Meno's Paradox

At this point in the dialogue, Meno is ready to throw in the towel: all his attempts to define virtue have failed, and Socrates has already admitted that he cannot define it either[21]; how are they to make progress? In contrast, Socrates is more optimistic about their prospects: although he claims to be as ignorant about the nature of virtue as Meno, perhaps they will find the answer they seek if they inquire diligently into the matter together. But Meno is not so sure, for he believes that the very idea of such an inquiry is incoherent:

> But how will you look for something when you don't in the least know what it is? How on earth are you going to set up something you don't know as the object of your search? To put it another way, even if you come right up against it, how will you know that what you have found is the thing you didn't know[22]?

Socrates replies:

> I know what you mean. Do you realise that what you are bringing up is the trick argument that a man cannot try to discover either what he knows or what he does not know? He would not seek what he knows, for since he knows it there is no need of the inquiry, nor what he does not know, for in that case he does not even know what he is to look for[23].

This is the well known Meno paradox. In this section, I suggest that although the paradox may not be that difficult to resolve, its resolution leaves some interesting issues about the role of hope in inquiry in its wake. These issues are taken up in the remainder of this chapter, and in chapter 2.

Let us take Socrates' speech first, as it is the clearest; it may be set out as follows[24]:

1. For any x, one either knows, or does not know, x.
2. If one knows x, one cannot inquire into x.
3. If one does not know x, one cannot inquire into x.
4. *Therefore*, whether or not one knows x, one cannot inquire into x.

[20] See, for example, C. S. Peirce, *Collected Papers* 6.4 and 6.273.

[21] At *Meno* 71b. In chapter 3, I argue that Meno is heavily dependent upon other people for his opinions, having no capacity for critical inquiry himself. So, he is bound to be discouraged by Socrates' confession of ignorance. See pp106-108 and p116.

[22] *Ibid.* 80d.

[23] *Ibid.* 80e. The argument is obviously familiar to Socrates. This suggests that Meno may have picked it up from someone else - further evidence of his unwillingness or inability to play a constructive role in a genuine inquiry.

Meno's speech is best seen as raising doubts about the possibility of inquiry that tend to support the third premise of the above argument, for in it he implies that:

I. If one does not know x, one cannot set it up as the object of one's search.

II. If one does not know x, one will not be able to recognise it, even if one happens upon it.

The former states that one cannot so much as *frame* an inquiry properly, if one is completely ignorant of its object. The latter states that, even if one could embark upon such an inquiry, it would be impossible to complete it in a satisfactory manner. But how plausible is any of this? Inquiry certainly *seems* to be possible; indeed, I believe I am engaged in philosophical inquiry at this very moment.

Before we can assess the above claims, there are a couple of issues we need to address. Firstly, we should say something about the nature of inquiry itself. The following rough and ready definition seems to capture most of the relevant intuitions: *inquiry is the directed, intentional search for knowledge or understanding which one lacks*[25]. According to this claim, inquiry is goal directed, the aim being to amass knowledge, or understanding. In order for such activity to be rational, one must presently lack whatever it is that one is searching for. And finally, the definition highlights the fact that inquiry is something we *do*, an *activity*. To get a better idea of what our rough definition of inquiry involves, think about what it implicitly excludes. The key word in the definition is "intentional". As this is the opposite of "unintentional", this implies that acquiring information accidentally, or in a passive manner, should not count as genuine inquiry. So, for example, simply seeing that an object has certain properties is not inquiry, nor is being told something by another person[26].

The second issue we need to address, before we can evaluate Meno's argument, is that there are at least three different ways of taking the premises of the argument, depending on the way in which the word *know* is being used:

[24] G. Fine gives the paradox in this form in "Inquiry in the *Meno*", p205.

[25] The definition is taken from G. Fine, *ibid*. n2.

[26] Again, these suggestions are taken from Fine, n2. Some qualification is needed though, for both sorts of cases can occur in the course of a genuine inquiry. Granted, it is not inquiry if I wander into a physics laboratory, and just happen to register the reading on a voltmeter. Yet, an observation of the very same meter reading by the physicist in charge of the experiment is, at least, part of her inquiry.

A. S knows x = S knows *that* x is the case

B. S knows x = S knows *something* about x

C. S knows x = S knows *everything* about x[27]

Since it is not clear from the dialogue which of the above ways of using the word *know* is intended, the discussion that follows will consider all three possibilities.

Let us reconsider the plausibility of Meno's argument against the possibility of rational inquiry in the light of all of the above. The table below sets out the various possibilities (each row represents a version of Meno's argument, according to the way in which we interpret its premises):

Table 1.1: Meno's Paradox

	Premise 1	**Premise 2**	**Premise 3**
Interpretation A	True	True	False
Interpretation B	True	False	?
Interpretation C	True	True	False

Not surprisingly, the first premise of the argument, being an instance of the law of the excluded middle, turns out to be true in all three cases[28]. The second premise is also true, as long as we assume that interpretation B is not intended. This follows from our definition of inquiry as a search for knowledge that one does not have: if one knows that x is the case, further inquiry into the matter is redundant; likewise if one already knows everything about x (supposing that idea makes sense) there is no reason to inquire further. In both of these cases, inquiry is not so much impossible as inappropriate. However, if one knows something (but not everything) about x, then inquiry into the x's unknown aspects is clearly possible, and may be a reasonable course of action.

The most contentious part of Meno's argument is obviously the third premise. If interpretation A is intended, this premise seems clearly false; not knowing that something is the case is not, in general, an obstacle to trying to find out *whether* it is the case. For

[27] One oddity is that it is not clear from the dialogue whether 'x' picks out a proposition, or an object, about which propositions can be known - all the more reason to distinguish carefully between the possible interpretations of Meno's argument.

[28] With interpretation C there is a temptation to read premise 1 as saying that one either knows everything about x or one knows *nothing* about x, but this would obviously be a mistake.

example, I don't know if there are any lentils in my cupboard right now, but it is surely a simple enough matter for me to find out (by going to the cupboard and looking, emptying it one item at a time if necessary). The third premise of the argument also looks false if we assume that interpretation C is intended. From the definition of inquiry, not knowing *everything* about the object of one's inquiry is a necessary condition for inquiry to be appropriate in the first place. More importantly, not knowing everything about the object of an inquiry does not generally preclude the possibility of starting and successfully concluding an inquiry into one of the object's unknown aspects, as long as it is possible to build upon what one already knows. For example, there are many things that scientists do not know about gravity; but it seems reasonable to suppose that, based upon what they already know about it and the other fundamental forces, they can devise cunning experiments in order to find out more.

It is only if interpretation B is intended that the third premise of his argument begins to look plausible: if one knows *nothing* whatsoever about x, the object of one's inquiry, how can one inquire into the nature of x? Significantly, the situation of radical ignorance about what it is one is looking for that this case suggests looks remarkably similar to the predicament Meno and Socrates find themselves in as they contemplate an inquiry into the nature of virtue. In addition, claims I and II, which I suggested are intended to support premise 3, also appear quite plausible if we assume that interpretation B is intended. For, taken together, these claims suggest that, without some knowledge of the object of one's inquiry at the outset, one will be able neither to begin nor complete the inquiry properly. The underlying thought appears to be that, as long as one is completely ignorant of what one is looking for, one will not be able to *control* the inquiry adequately. But, according to our definition, inquiry must be intelligently directed, or it is not inquiry at all. Hence, premise 3 appears to be true, on this reading; in which case the kind of inquiry into the nature of virtue that Socrates proposes is impossible[29].

I can think of two responses to this troubling conclusion. Firstly, is it really coherent to suppose that one might be contemplating some inquiry regarding x, without knowing anything at all about x? Do we not, at the very least have to suppose that one knows that x is the type of thing about which it is appropriate to inquire, or the kind of thing that others have considered inquiring into? Secondly, is it really the case that in order to inquire into x, one

[29] Or at least we aren't doing what we think we're doing when we engaged in such an inquiry. In particular, we are overestimating the degree of rational control we have over our inquiries.

needs to *know* something about it? Granted, if I wish to find out about the chemical properties of gold, but my mind contains nothing but falsehoods about it, my inquiry is not likely to be very successful. Errors have a habit of compounding themselves, until we find ourselves far from the truth. But does the starting point of such an inquiry have to be a state of full-blown knowledge if it is to succeed? Won't a somewhat less auspicious epistemic state - true belief, say - suffice to get things off on the right track, so that there is at least some chance of reaching the truth, as long as we conduct ourselves responsibly along the way? Both questions are reasonable, and the doubts they raise are, I believe, sufficient to deal with the immediate threat that the Meno paradox presents.

Many commentators have stopped at this point, believing that the issues that the paradox raises have thereby been dealt with. However, in my view, this is the point where things start to get interesting. As stated, the Meno paradox is probably[30] quite easy to resolve; but it leaves a number of questions in its wake that are interesting to think about, and difficult to answer. We shall be in a better position to appreciate these issues after we have examined Socrates' response to Meno's paradox.

1.2.5 The Theory of Recollection

Socrates' response to Meno's argument is the Theory of Recollection. In this section, I argue that this is best seen as part of a two tier response to the paradox Meno presents us with. The likely resolution of the paradox has already been discussed in the previous section. However, this resolution gives rise to further problems that the Theory of Recollection is supposed to resolve. I also argue that the Theory of Recollection is not something that we can know is true; the best that we can do is entertain it as a regulative *hope* in the course of our inquiries. At the end of the section, I suggest that, according to the *Meno*, there are two hopes that we need to entertain if the activity of inquiry is to seem coherent to us: that the universe is orderly; and that our intellectual faculties are attuned to that order. As we shall see in section 1.3, Peirce arrives at strikingly similar conclusions about the role of hope in inquiry.

Socrates is not perturbed by the paradox, and his description of it as a "trick" argument suggests that we should not allow it to trouble us either. When Meno asks him whether he thinks it is a good argument, and if not whether he can explain how it fails, he responds simply and confidently: no, it is not a good argument, and yes, he can explain how it fails.

[30] I don't claim that the above remarks constitute a definitive refutation of Meno's argument; just that they raise enough doubts about its soundness, for us not to have to worry about it in what follows.

All of this is very much in line with my remarks in the previous section. But Socrates' confidence begins to look a little odd when set against what he believes is necessary to refute Meno's argument. He pauses solemnly, before recounting to Meno something he has heard from "men and women who understand the truths of religion":

> ... the soul since it is immortal and has been born many times, and has seen all things both here and in the other world, has learned everything that is. So we need not be surprised if it can recall the knowledge of virtue or anything else which, as we see it once possessed. All nature is akin, and the soul has learned everything, so that when a man has recalled a single piece of knowledge - *learned* it in ordinary language - there is no reason why he should not find out all the rest, if he keeps a stout heart and does not grow weary of the search; for seeking and learning are in fact nothing but recollection.
>
> We ought then not to be led astray by the contentious argument you quoted. It would make us lazy and is music to the ears of weaklings[31].

Is this dose of industrial-strength metaphysics really the only cure for the ills that Meno's paradox threatens us with? If so, it seems rather blasé to dismiss the challenge as nothing more than a "trick" argument - how many other tricks demand such a drastic response? Furthermore, it seems odd that Socrates has so little evidence for such an extravagant claim; it is just something that he has heard from people who know about such things, and he seems to expect Meno to take the story on trust as well[32]. It would be nice if we could just ignore the theory as an aberration on Socrates' part, but the circumstances - he speaks in a lowered voice, as if imparting a great secret - indicate that he intends it to be taken with the utmost seriousness.

So, what are we to make of the Theory of Recollection, and its relationship to the paradox that precedes it in the dialogue[33]? Recall how we sought to defuse the paradox in the previous section: it is not true that knowledge of X is necessary in order to inquire about X; true beliefs about X will usually suffice. Of course, it is important that at least some of the beliefs we hold about X at the start of the inquiry are indeed true. If we start out with nothing but falsehoods, it is hard to see how inquiry can succeed. This point holds even though, from the inquirer's point of view, the beliefs that are, as a matter of fact true, are indistinguishable from the beliefs that future inquiry will eventually reveal as false. All of this implies that the above response to Meno's paradox only works as long as we have a tendency to favour true

[31] *Meno* 80c-d.

[32] As I explain below, however, Socrates' commitment to the Theory of Recollection is somewhat qualified.

over false beliefs[34]. But how can this assumption be justified? What could account for such a capacity, if, indeed, we really do have it? One possibility is that we are already in possession of the truth - always have been, and always will - we just have to work hard to remember it sometimes. In other words, the Theory of Recollection is introduced to account for our tendency to hit upon true rather than false beliefs in the course of our inquiries, a tendency that is invoked to deal with Meno's paradox. Gail Fine summarises this position as follows:

> ... the Theory of Recollection is introduced, not as a direct reply to the paradox... but to explain certain facts assumed in [the] reply. For example, [the] reply assumes that in inquiring we tend to favour true over false beliefs. Plato believes that this remarkable tendency cannot be a brute fact, but requires further explanation; the best such explanation, in his view, is the Theory of Recollection[35]

In what's left of this section, I want to raise some questions about this interpretation, and in particular about the adequacy of what we might call the two-tiered response to Meno's paradox it imputes to Socrates. In doing this, I am preparing the ground for the discussion of Peirce's theory of inquiry in section 1.3, and the consideration of the role of hope in inquiry that occupies chapter 2. The two tier response that Fine claims to find in the dialogue is shown in Table 1.2.

Table 1.2: The Theory of Recollection and Meno's Paradox

	Problem	**Solution**
Tier 1	How is inquiry from a position of initial ignorance possible?	It is possible as long as some of our initial beliefs about the object of inquiry are, in fact, *true*.
Tier 2	How is it that a enough of our initial beliefs are true to make possible stable, successful inquiry?	Theory of Recollection: The truth of the initial beliefs may be accounted for by postulating that we already know all of the relevant truths, and have just forgotten many of them.

If we are after reassurance that genuine inquiry is possible then, everything hinges on the claim in the bottom right hand corner of the above table. But how much reliance are we entitled to place on the Theory of Recollection? Socrates tells Meno that he has heard the

[33] The interpretation that follows is put forward by Fine in "Inquiry in the *Meno*". See especially pp213-15.

[34] In section 1.3 of this chapter, we shall see that this is a theme that Peirce makes much of.

Theory of Recollection espoused by "men and women who understand the truths of religion"[36]. Having heard the story from people he believes to be reliable, it might be thought that Socrates would be prepared to claim knowledge in this instance. There *are* a couple of occasions in the dialogue where Socrates expresses a firm belief in the Theory of Recollection. For example, he describes what the priests and priestesses have told him as "something true, I think, and fine", and a little later on describes himself as "convinced of its truth" [37]. But there is also an odd note of caution in his treatment of the doctrine, best illustrated by his comment after the dialogue with Meno's slave boy:

> I shouldn't like to take my oath on the whole story [the Theory of Recollection], but one thing I am ready to fight for as long as I can, in word and act: that is, that we shall be better, braver and more active men if we believe it right to look for what we don't know than if we believe there is no point in looking because what we don't know we can never discover[38].

In a way this is an unremarkable passage. Socrates is simply saying that he isn't prepared to swear that the Theory of Recollection is true, isn't, we might say, absolutely certain of it. Given the heavy metaphysical commitment the theory involves this is a rational attitude to take. But there is also a suggestion that, setting aside the difficulty of rationally evaluating the claim that the priests and priestesses make, the Theory of Recollection is still something we ought to entertain because it is the only thing that offers us any *hope* that the activity of inquiry that we find so irresistible is not ultimately misguided. In short, there is a suggestion that, at least in the absence of any evidence to the contrary, we are entitled to entertain the Theory of Recollection on pragmatic grounds as a kind of regulative hope[39]. This interpretation of what is going on makes sense of Socrates' earlier insistence that, as long as a man has recalled a single piece of knowledge, there is no reason why he should not learn all of the rest as long as he "keeps a stout heart and does not grow weary of the search"[40]. Almost immediately after this, Socrates criticises Meno's paradox on the grounds that it is "music to the ear of weaklings", something that would make us lazy if we accepted it.

[35] G. Fine, "Inquiry in the *Meno*", p213.

[36] *Meno*, 81a.

[37] *Ibid* 81a, 81e.

[38] *Ibid* 86b.

[39] Another possibility is that the Theory of Recollection is offered as an inference to the best explanation. Given that inquiry is possible, then something like the Theory of Recollection must be true. Nowadays, of course, most of us do *not* regard the Theory of Recollection as an especially good explanation, preferring theories of innate knowledge in its place. Still this is just an uncharitable response to our philosophical ancestors. The main flaw with this suggestion is that it is simply not available in these circumstances, where the very possibility of genuine inquiry is what's in doubt.

[40] *Ibid* 81d.

In conclusion, the Theory of Recollection is not a direct response to Meno's paradox, but is rather something that has to be invoked to ameliorate some of the problems the most obvious solution to the paradox leaves behind. In this, I am in agreement with Gail Fine. However, it is important to realise that the theory has a rather special function. It is not something we accept on the basis of the evidence that is available to us - given what it claims, how could it be? Instead, it is a conviction that we somehow have to cultivate, perhaps despite our misgivings, if our intellectual life is to have any coherence. And in Socrates' mouth, it is a rallying cry against intellectual apathy and despair.

Before leaving this section, it is worth looking again at the speech that introduces the Theory of Recollection:

> ... the soul since it is immortal and has been born many times, and has seen all things both here and in the other world, has learned everything that is. So we need not be surprised if it can recall the knowledge of virtue or anything else which, as we see it once possessed. *All nature is akin*, and the soul has learned everything, so that when a man has recalled a single piece of knowledge - *learned* it in ordinary language - there is no reason why he should not find out all the rest, if he keeps a stout heart and does not grow weary of the search; for seeking and learning are in fact nothing but recollection.

This time I have italicised the phrase *all nature is akin*; what does this mean? The most natural reading, I think, is that the universe contains a high degree of regularity and order. Aspects of the world in which we dwell are never entirely unique, but resemble one another in various respects, to greater or lesser degrees; just as members of the same family resemble one another. Although, Socrates seems to state that this is the case unequivocally, given our previous remarks and the context, it is likely that the kinship of all nature is something that we can only ever hope is the case, not something that can be established conclusively.

If this is correct, the regulative hope that underlies the possibility of rational inquiry has a dual aspect: it is a hope in the essential orderliness of the object of our inquiry; and - in the form of the Theory of Recollection itself - it is a hope in our ability to discern that order[41]. As I explain in section 1.3, it is just these hopes that play such a prominent part in Peirce's theory of inquiry.

[41] Not that the Theory of Recollection is the only possible way of accounting for the apparent attunement of our intellects to the order in the universe. For example, as we shall see in section 1.3, Peirce flirted with evolutionary explanations for this attunement.

1.2.6 The Slave Boy Dialogue

Meno asks Socrates to provide whatever evidence he can for the truth of the Theory of Recollection. In response, Socrates offers to demonstrate the theory in action by helping one of Meno's slaves to "recollect" an elementary geometrical truth. The slave boy is given no name in the dialogue, but I shall call him*Nemo*. Socrates challenges Nemo to determine the length of the side of a square having twice the area of a given square. The dialogue within a dialogue that ensues is significant, because it has a bearing on all three of the principal claims of this chapter. The best way to show this is by raising a number of worries about the adequacy of the dialogue as a proof of the Theory of Recollection.

The first worry is this: Meno's apparent concern is with the possibility of inquiry when nothing whatsoever is known about the object of inquiry, yet Socrates clearly*does* know the solution to the geometrical problem he sets Meno's slave. The whole episode therefore appears to be an example of *teaching* rather than inquiry. If that is correct, how can it address Meno's worry about the possibility of inquiry[42]?

To resolve this difficulty, we must break down the rigid distinction between teaching and inquiry that it presupposes. One way of doing this is to note that, in his conversation with Socrates, the slave boy does not just passively accept everything that Socrates tells him. As a matter of fact, Socrates tells Nemo very little; most of the dialogue is conducted via a series of questions set by Socrates that the slave does his best to answer. The series of questions is carefully chosen to lead Nemo to the solution of the original geometrical problem via a series of short steps.

Another way of putting the above point: the objection that the slave boy dialogue is an example of teaching rather than inquiry would be valid if Socrates simply *told* Nemo the answer to the geometrical problem. For, when one is engaged in genuine inquiry, *ex hypothesi*, there is no one around to tell one the answer to the question being investigated; somehow, one has discover the answer for oneself. In a sense, this is just what Nemo does, albeit with guidance from Socrates.

At this point, a second worry arises: the slave boy may not be entirely passive in the dialogue, but surely he receives *too much* help from Socrates. For, whilst it may be an

[42] An alert reader of the dialogue may have anticipated trouble of this sort earlier on, when Socrates refers to both seeking and learning as forms of recollection (*Meno* 81d).

exaggeration to say that Socrates *tells* Nemo the answer to the geometrical problem, examination of the dialogue reveals many instances of Socrates apparently putting words into the slave boy's mouth by asking *leading* questions. Here are just a few examples[43]:

> It has all these four sides equal?
>
> And these lines which go through the middle of it are also equal?
>
> But since it is two feet this way also, does it not become twice two feet?
>
> But does it not contain these four squares, each equal to the original four feet one?
>
> Won't it be four times as big?

One way out of this difficulty is to highlight what has been called the 'say what you think' constraint. Although many of Socrates' questions suggest one answer rather than another, the effect of these leading questions is, at least partly, neutralised by Socrates insistence, on several occasions, that Nemo always respond with his own opinions:

> Always answer what you think[44].
>
> Be ready to catch me if I give him any instruction or explanation instead of simply interrogating him on his own opinions[45].
>
> What do you think Meno? Has he answered with any opinions that were not his own[46]?

Whilst this reply goes some way towards answering the above objection, I do not think it is the whole story. However, drawing attention to the operation of the 'say what you think' principle within the dialogue does point to a more satisfactory response. Suppose we just accept that, at many points in the dialogue, Socrates all but tells Nemo the steps of the geometrical proof. Still he insists that the slave boy is engaged in a genuine effort to recollect the solution to the geometrical problem. This means that the slave boy must still be inquiring even *after* being led so far along the path by Socrates. In that case, inquiry must, in Socrates' view, be something that happens over and above responding to particular questions. What this additional element might be is indicated by Socrates insistence that the boy answers honestly, with his own opinions. This suggests that Nemo must ultimately be able to demonstrate something like a*personal commitment* to the truth of what he says in the

[43] All from *Meno*, 82c-83b
[44] *Ibid.* 83d
[45] *Ibid.* 84d
[46] *Ibid.* 85b

dialogue. In other words, even if Socrates is asking leading questions, it is not enough for Nemo to blindly *follow* that lead. There is no place in genuine inquiry for this kind of deference to authority[47]; genuine inquiry requires *autonomy*. How such autonomy is manifested is not clear at this stage, but I suggest that it involves a readiness to defend our opinions against all manner of objections, many of which will be unexpected. Obviously, we cannot do this if we are limited to repeating what other people have told us; we can only do it if we make an effort to *make sense* of what other people say[48]. This conclusion is supported by the fact that Socrates' does not attribute knowledge to the slave boy at the end of the dialogue, at least not in the full blooded sense that he regards as the true end of inquiry. Instead, we hear him tell Meno that:

> At present. these opinions, being newly aroused, have a dream-like quality. But if the same questions are put to him on many occasions and in different ways, you can see that in the end he will have a knowledge on the subject as accurate as anybody's[49].

Note that, from a modern perspective, there is a puzzle here: Nemo now has a justified true belief about a particular geometrical topic, but this is still not the end of the matter as far as Socrates is concerned. This suggests, among other things that *episteme*, the term used in the dialogue to describe the goal of inquiry, should not, as it usually is, be translated as "knowledge", since the slave boy certainly has *that* at the end of his dialogue with Socrates. I return to this topic, and to the nature of the intellectual labour required to convert true opinions (even *justified* true opinions) to *episteme* in chapter 3[50].

Meanwhile, let us consider a third and final worry about the slave boy dialogue. Remember, it is supposed to be an example of a successful inquiry that starts out from a state of complete ignorance; as such, it is intended as a response to Meno's scepticism about the possibility of conducting an inquiry into the nature of virtue. The puzzle is that the two inquiries are not analogous: in the original, abortive, inquiry, *both* participants are ignorant, whereas in the slave boy dialogue, Socrates *knows* the solution to the geometrical problem. Hence, even if we set aside our worry that Socrates is blurring the distinction between learning and inquiry, and turn a blind eye to his leading questions, there is still cause for concern. For it may be argued that Nemo is only able to arrive at the solution to the geometrical problem because Socrates leads him to it via a carefully chosen sequence of

[47] The fact that this part of the dialogue is with a *slave* boy is thus significant, on my interpretation.
[48] And of the "testimony of the senses" in general, we might add. I return to this issue in my discussion of, what I call, *non-minimal comprehension*, in chapter 3.
[49] Plato, *Meno*, 85c.

questions[51]. But surely Socrates is only able to choose his questions so carefully because he already knows the solution to the problem he sets the slave.

This objection draws our attention to the fact that progress in any inquiry usually depends upon asking the right *questions*. This being the case, I think we must concede that Socrates' dialogue with the slave boy is not, in itself, adequate as a response to Meno's worry about the possibility of inquiring into the nature of virtue[52]. The slave boy, we might say, is lucky to have Socrates around feeding him questions that will further his inquiry at the appropriate time, but in most of our inquiries we are not so fortunate.

Does this mean, as Meno suggests, that those inquiries are doomed from the outset? Not necessarily: as long as we have someone or *something* to guide us, we can reasonably expect to make progress. In the case of most of our inquiries, I suggest that what guides us in framing our questions is the subject matter of the inquiry itself, combined with our existing knowledge. Attention to the content of our experience in the course of an inquiry *suggests*, as it were, questions that we ought to ask ourselves. Of course, there is no *guarantee* that answering these questions will enable us to successfully complete our inquiry, but - to return to the theme of the previous section - it is surely reasonable to hope that it will do so. The only alternative to entertaining this hope - that in a given context of inquiry, the questions that we are inclined to ask tend to be ones that will further the inquiry - is a kind of intellectual despair that most of us will find unappealing.

1.2.7 The Acquisition of Virtue (II)

At this point in the dialogue, Meno has apparently accepted that a joint inquiry into the nature of virtue is possible. It comes as something of a surprise therefore when, upon being invited by Socrates to resume the quest, he replies:

> All the same , I would rather consider the question as I put it at the beginning, and hear your views on it; that is are we to pursue virtue as something that can be taught, or do men have it as a gift of nature or how[53]?

[50] See especially the discussion of non-minimal comprehension, in section 3.4.

[51] Note that this would be a valid concern even if none of Socrates' questions were 'leading' ones.

[52] This does not necessarily mean that the dialogue is flawed, for the disanalogy between the two inquiries may be deliberate. The fact that Meno cheerfully accepts what Socrates tells him without noticing this rather obvious disanalogy may be designed to show that Meno is an intellectual slave, incapable of genuine inquiry.

[53] *Ibid.* 86c

Even more surprising is the fact that Socrates agrees to this request, having previously argued that such an inquiry, prior to understanding the essence of virtue, is not possible.

I shall not comment on the section of the dialogue that follows, as it is not directly relevant to the subject of this thesis. More relevant is the fact that Meno, having previously agreed with Socrates about the "priority of definition" reverts to his original question about the source of virtue without making a serious attempt to define it. True, this could just be an innocent lapse of memory, but, in the light of our discussion of the slave boy dialogue, I think that another interpretation is possible. Whatever it is that Nemo must do in order to arrive at a complete understanding of the solution to the geometrical problem, it appears that Meno has not been able to do with regard to his discussions with Socrates about the nature of virtue. Although he *appears* to have been persuaded by Socrates' insistence that we need to know the definition of something before we can inquire into other aspects of it, it now looks like he was just paying lip service to the requirement. In the terms of the previous section, Meno has deferred to Socrates' authority on the topic, and made no serious effort to exercise his intellectual autonomy and understand what he has been told.

1.2.8 True Opinion vs Knowledge

The final section of the dialogue is interesting for what it suggests about the nature of understanding. I shall keep my discussion of it brief, as the theme is developed more fully in chapter 3.

As noted in section 1.2.6, inspection of the slave boy dialogue reveals that even at its conclusion, Nemo's epistemic state falls short of the ideal, as Socrates conceives it. And this despite the fact that, to a modern reader, Nemo *knows* the solution to the geometrical problem, at least if knowledge has anything to do with justified true belief[54]. The final section of the dialogue provides a clue as to what more Socrates is asking of the boy. His complaint at the end of the slave boy dialogue is that the true opinions in his interlocutor's mind are newly aroused, and therefore have a kind of dreamlike quality to them. It is characteristic of dreams that they tend to evaporate upon contact with reality (think of all the times you have struggled to remember later in the day the tiniest fragment of a dream that seemed so arrestingly vivid upon waking). This suggests that the ideal epistemic state that inquiry aims at has a sort of stability or robustness to it.

[54] Abstracting, for the sake of expositional clarity, from Gettier type considerations.

This suggestion is amplified in the final section of the dialogue, where Socrates attempts to formulate a distinction between true belief (*orthe doxa*) and *episteme*, which, I have suggested, should be translated as "understanding", rather than "knowledge"[55]. Previously, it has been established to both Socrates' and Meno's satisfaction that both true belief and *episteme* are equally good as bases for action[56]. To see this, suppose, believing my little boat to be leak proof, I attempt to cross the icy fjord that separates my isolated cottage from the nearest village. As long as the boat really is leak free, I shall be able to carry out my plan. As far as success in action is concerned, it does not seem to matter how I arrived at the belief - careful checking of the hull, and receiving reassurance in the form of a vision stand on an equal footing in this respect. But if this is the case, Meno wonders, why do we value *episteme* so much more highly than mere true belief?

Socrates' answer recalls his earlier description of Nemo's "dreamlike" new opinions. True opinions are, he says, like the legendary statues of Daedalus, which, though valuable, would run away unless they were tied in place. True beliefs are indeed useful as a basis for action - just as useful as *episteme* - but the problem is that - unlike *episteme* - they cannot be relied upon to stick around:

> True opinions are a fine thing and do all sorts of good so long as they stay in their place; but they will not stay long. They run away from a man's mind, so they are not worth much until you tether them by working out the reason. That process, my dear Meno, is recollection, as we agreed earlier. Once they are tied down they become *episteme*, and are stable[57].

It is the relative stability of *episteme*, that, according to Socrates, accounts for its being more highly esteemed than true opinion. If, as I have suggested in sections 1.2.1 and 1.2.6, *episteme* is best translated as "understanding", this implies that understanding too has a distinctive kind of stability. I return to this issue in chapters 3, 4 and 5 of the thesis.

1.3 Peirce's Theory of Inquiry

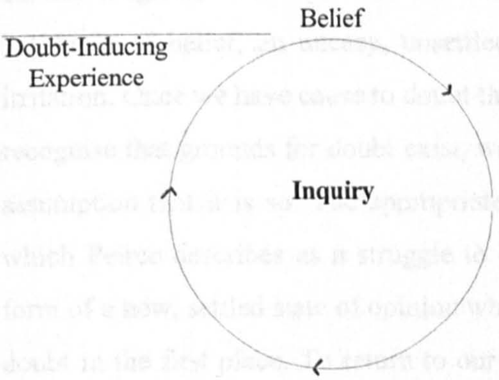
We saw in the previous section that, according to the *Meno*, the goal of inquiry is understanding. The dialogue does not underestimate the difficulty of attaining this goal, but suggests that we shall fare better in our inquiries if we adopt an optimistic outlook. The

[55] See sections 1.2.1 and 1.2.6.
[56] *Ibid.* 96e-97c
[57] *Ibid.* 98a

Theory of Recollection is designed to promote this. Neither of these key themes is heavily emphasised in the *Meno*, but, arguably, the suggestion that the goal of inquiry is understanding receives rather more attention[58] than the idea that hope has an important role to play in the regulation of inquiry. This situation is more or less reversed in Peirce's writings on the nature of inquiry: understanding is treated only implicitly, whereas there are many tantalising remarks about the role that fundamental hopes play in our inquiries. Hence, the brief description of Peirce's theory of inquiry[59] that I offer in this section is designed to complement the previous section's analysis of the *Meno*.

The most important thing to bear in mind when reading Peirce's work on this topic is that he regards inquiry as a *functional* part of a larger whole, something which, for want of a better expression, might be called "the scientific way of life"[60]. The following diagram is an attempt to explain what this means:

Figure 1.1: Peirce's Model of Inquiry



According to Peirce, inquiry begins and ends with belief, a state of mind with the following properties[61]:

1. Belief has a *phenomenology* that distinguishes it from other propositional attitudes.

[58] Insofar as Socrates takes pains to distinguish between *episteme* and true opinion. I have said relatively little about this material in this chapter because part of chapter 3 is devoted to exploring it.

[59] Peirce's theory of inquiry, and the remarks about hope that he makes in connection with it, are scattered throughout his collected works. The account that follows draws on *Collected Papers*, 5.171 to 5.173, 5.591, 7.220 and 5.160. Volume 7, book II of Peirce's *Collected Papers* also contains relevant material.

[60] Not that it is confined to professional scientists; rather, it is present wherever individuals are making an intelligent effort to understand the world and their place within it. Peirce does tend to concentrate on the special case of scientific inquiry though. However, since we have already determined that inquiry in the *Meno* is a search for real essences, and is therefore "scientific" in a broad sense, this ought not to bother us.

2. Belief is what we are prepared to *act* upon, if the circumstances are appropriate.
3. Belief is a "calm and satisfactory" state; once in it, we are not motivated to inquire further.

For example, believing that the ice on the lake is thick enough to hold my weight (arguably) feels different from wondering whether it is. Because I have such a belief, I am prepared to put on my skates and venture onto the frozen lake, if the desire to go ice skating arises. Finally, unless new information becomes available, I am not inclined to inquire further into whether it is safe to skate on the lake; as far as I am concerned, the matter is settled. Belief then, is a kind of *equilibrium* state, which enables us to have successful - that is, desire satisfying - interactions with the world and other people. However, as the above diagram illustrates, the equilibrium in question is *dynamic*: the world constantly presents us with reasons to revise our existing beliefs, to reflect upon our habitual patterns of activity, and to modify them in line with fresh information.

In the diagram, such episodes are labelled doubt-inducing experiences. Doubt is the antithesis of belief, an uneasy, unsettled state of mind that we experience as a kind of irritation. Once we have cause to doubt that something is the case, and are rational enough to recognise that grounds for doubt exist, we are no longer prepared to plan our actions on the assumption that it is so. The appropriate response to this problematic situation is inquiry, which Peirce describes as a struggle to overcome doubt and to restore equilibrium in the form of a new, settled state of opinion which takes on board the information that induced the doubt in the first place. To return to our skating example, it may be that as I approach the frozen lake, I notice a thin crack in the ice. My confidence that the lake is safe to skate on is shaken, and I am no longer prepared to act on that assumption. Most likely I shall conduct further tests - throw rocks onto the ice, and so forth - the results of which will either confirm or disconfirm my original belief. So, according to Peirce, inquiry is an intelligent response to a state of affairs which has become problematic; it is an intentional activity which aims to bring about a new state of cognitive equilibrium by eliminating the irritation of doubt[62].

So far, I have presented a rather crude sketch of the general outlines of Peirce's theory. In a moment, I shall home in on some details that are especially interesting, from our point of

[61] As described in "The Fixation of Belief", section III.

[62] Note that Peirce is careful not to insist that the object of inquiry is to install a state of *true* belief. Such an additional requirement is redundant since: (i) To believe something and to believe that it is

view. Before that, it is worth pausing to note some interesting parallels between the above material, and the model of inquiry that is implicit in the*Meno*. At the beginning of Plato's dialogue, Meno's mind is made up; he already has fixed opinions about the nature of virtue, and he is quite prepared to act upon them - to speak confidently about virtue to large assemblies, and so forth. He remains untroubled by doubt until Socrates forces him to examine his beliefs carefully. Socrates' patient questioning uncovers many inconsistencies, and eventually reduces Meno to *aporia*, an unpleasant state of intellectual paralysis. According to Peirce's theory, this is the point at which genuine inquiry into the nature of virtue becomes possible, even necessary. Sure enough, this is just what Socrates proposes, only to encounter reluctance on the part of Meno. Finally, it is worth registering that Socrates, like Peirce, believes that inquiry is an arduous activity, a struggle, but that success is possible if we persevere. No doubt the difficulty of conducting a proper inquiry into the nature of virtue partly accounts for Meno's inertia.

So, according to Peirce, the catalyst for scientific inquiry is a doubt-inducing experience. Since a working scientist is conditioned to see the workings of the physical universe as the inevitable outcome of regular laws, she comes to form expectations about the future course of events, based on her present understanding of those laws. If and when these expectations are not met, she will most likely experience surprise, and will begin to doubt whether her present understanding of the laws of nature is adequate. At this point, inquiry is necessary, in order to determine the reason for the occurrence of the anomaly. Such an inquiry aims to *explain* the puzzling phenomenon, by identifying some proposition, or set of propositions which, if true, would render the observed phenomenon unsurprising after all[63, 64].

To go into a little more detail, scientific inquiry, according to Peirce, has three distinct components. The first of these, the most interesting, as far as we are concerned, he calls *abduction*, or, occasionally, retroduction. Abduction is the adoption of a hypothesis, likely in itself, which renders the surprising facts likely also. I shall have more to say about abduction very shortly. The second component of scientific inquiry, *deduction*, will be more familiar to most readers. Having adopted a working hypothesis, we trace out its necessary and probable experiential consequences. After performing this task we are in a position to design

true are the same thing; and (ii) "Only believe something if it is true" is advice that it is impossible to follow, for it is not possible to discover that I have failed to follow it while still holding the belief.

[63] Again, this accords with our characterisation of Socratic inquiry: the search for a real definition of virtue, say, that accounts for our pre-theoretical intuitions.

[64] For more on explanation in science (including a brief summary of Hempel's theory of scientific explanation, which Peirce's ideas prefigure), see section 5.4.

experiments capable of testing our original hypothesis, the third, and final, stage of scientific inquiry. This Peirce calls, somewhat confusingly, *induction*. Once the experimental results are in, we are in a position to reject, or accept, pending evidence to the contrary, our original hypothesis; or perhaps we will be forced to formulate further tests, if the results are inconclusive. Of course, this description of the process of inquiry is very crude. In particular, it makes it sound as if the three stages are logically independent, each issuing from the preceding one in a rigidly linear fashion. This is certainly simplistic: the hypotheses one chooses to explore, for instance, are, in real life, conditioned by what is presently amenable to empirical testing. Nevertheless, the three aspects of inquiry have a certain degree of autonomy which allows us to examine them individually. This is just as well since I now propose to focus exclusively on abduction.

There are two reasons for giving preferential treatment to abduction. Firstly, as argued in the previous section, the quest for a definition of virtue in the *Meno* is a search for an explanatory hypothesis, a description of virtue's essence that accounts for our pre-theoretical intuitions (about who qualifies as virtuous, and so forth). According to Peirce, abduction is the stage of inquiry concerned with the generation of possible explanations. Secondly, deduction and induction are familiar and well understood aspects of inquiry. In addition, once we have an explanatory hypothesis in mind, it is a relatively simple matter to deduce its experiential consequences, and to come up with critical empirical tests. In comparison, abduction appears quite mysterious. How do scientists set about generating explanatory hypotheses? The thought processes involved can scarcely be entirely random, yet it is just as difficult to imagine that they are very systematic. Also, abduction is arguably the most important element in inquiry; for without a steady flow of reasonable, explanatory hypotheses, the whole process grinds to a halt. As Peirce remarks, abduction is "the only logical operation which introduces any new idea"[65] into the process of rational inquiry. All of this suggests that, if rational inquiry is impossible, as Meno seems to think, the difficulty is quite likely to be traceable to its abductive element.

What Peirce says about abduction is complex and interesting, though only a small part of it is directly relevant to our inquiry. Broadly speaking, he claims, it is possible to identify three factors that govern our choice of a hypothesis for experimental testing:

1. It must lend itself to experimental testing.

[65] C. S. Peirce, *Collected Papers*, 5.171.

2. It must explain the surprising facts.
3. It must be economical.

The first of these is a special application of Peirce's pragmatic maxim, according to which the significance of a proposition is to be understood in terms of the difference its truth would make in experience[66]. Whatever its merits as a general principle, this is a sensible maxim for a working scientist to adopt; such an individual has nothing to gain from framing explanatory hypotheses that are, in principle, impossible to test experimentally[67]. The second item on the above list simply states that, whichever hypothesis we choose, it must be capable of explaining the surprising facts that precipitated the inquiry. That is, says Peirce, it must be possible to deduce the puzzling facts as necessary or likely consequences of the truth of the hypothesis. The third factor is not self-evident, but, upon reflection, is just as necessary as the first two. There are many, possibly infinitely many, empirical hypotheses that explain the facts equally well, in Peirce's sense. So, given that the time and intellectual resources at our disposal are finite, some method must be found of rationing these scarce resources, in order to maximise our chances of selecting the correct hypothesis. Hence, there is a need for a theory that deals with the economics of hypothesis selection. Peirce identifies three purely economic factors that influence hypothesis choice, the two main ones being *cost*, and *value*[68].

Cost includes such things as the amount of time testing the hypothesis will take, and the material and intellectual resources that this is likely to require. Less obviously, it includes the likely effect of the hypothesis on our other beliefs. Since very few, if any, of our beliefs are completely isolated from the rest, the adoption of a new hypothesis has ramifications throughout the network. The process of adjustment to a new equilibrium which this necessitates is itself costly in terms of time and mental resources, and these hidden costs must be taken into account. The rationality of adopting a given working hypothesis depends partly upon the total cost that course of action entails. Likewise, devoting time and energy to exploring a hypothesis is rational to the extent that the hypothesis is of value. Several factors

[66] The classic statement of the maxim is in "How to Make Our Ideas Clear", section II.

[67] In practice things are not so clear cut. When first formulated, the hypothesis that all of the known sub-atomic particles, including protons, neutrons, and electrons are made of just three "flavours" of quarks was widely regarded as immune from experimental refutation, as, according to the theory, an isolated quark was impossible to observe. Still, the idea was not killed outright, as it enabled theoreticians to classify particles in an extremely elegant manner. In the meantime, a large amount of indirect experimental evidence in support of the quark hypothesis has accrued. For an excellent, very readable account of this episode, see R. P. Crease and C. C. Mann, *The Second Creation*, chapters. 14 and 15.

[68] I discuss these in the next paragraph, where I gerrymander Peirce's classification slightly, for the sake of simplicity.

contribute to the value of a hypothesis, but three, simplicity, breadth, and plausibility, are especially important. Clearly, the simpler the hypothesis, and the greater the range of its predicted observational consequences, the more valuable it is. Similarly, the more likely, in our estimation, a hypothesis is to be true[69], the more valuable it is, other things being equal; and, conversely, there is little point in testing a hypothesis if it just isn't very likely to be true[70].

Assessing how likely a given hypothesis is to be true is a black art, however. Some scientists are evidently much better at it than others, but it is difficult to say why. Peirce suggests just two ways of accounting for this ability, which we all have to a greater or lesser extent, to home in on plausible explanatory hypotheses. Firstly, we tend to regard a hypothesis as plausible if it fits in with our preconceived ideas about the way the phenomenon we are studying works. Selecting hypotheses for further exploration on this basis seems reasonable; if nothing else, following this advice will keep down the costs of inquiry, since it counsels a kind of intellectual conservatism. Secondly, Peirce believes that, in the course of their investigations, scientists develop what he calls an *instinct* for the truth. This hardly seems very explanatory, but Peirce is quite insistent on this point: "...the existence of a natural instinct for truth is, after all, the sheet-anchor of science"[71].

Why does Peirce believe that it is necessary to postulate the existence of this instinct, and what does he say to support his hypothesis? Answering these questions brings us into contact with some of Peirce's most provocative, but also frustratingly schematic, ideas about the nature of scientific inquiry. His main argument for this conclusion uses a process of elimination. Firstly, abductive logic never *compels* us to take seriously any given explanatory hypothesis. As evidence for this claim, Peirce points out that scientists are rarely able to give reasons for their best guesses; more often than not, the idea just comes to them[72]. This contrasts with the case of deductive logic: having accepted a set of premises, we must, on pain of inconsistency, endorse the propositions that they entail. Secondly, hitting upon plausible explanatory hypotheses cannot be a matter of chance either, for the number of possibilities is just too vast:

[69] In this, and the next, paragraph, I use the phrase "likely to be true" in the same sense as "plausible", and do not mean to suggest that probabilities are literally assigned to hypotheses before testing, or revised during testing. Some approaches to inductive logic (e.g. Bayesian ones) deal in such probabilities, but we do not need to, for the purposes of this thesis.

[70] Unless it is *very* cheap to test.

[71] C. S. Peirce, *Collected Papers*, 7.220.

[72] C. S. Peirce, *Collected Papers*, 5.173.

> A physicist comes across some new phenomenon in his laboratory. How does he know but that the conjunctions of the planets have something to do with it or that it is not perhaps because the dowager empress of China has at that time a year ago happened to pronounce some word of mystical power[73].

Peirce encourages us to exercise our imaginations, and to think of the "trillions of trillions" of hypotheses that might be entertained, only one of which is true. Yet, as history shows, scientists seem to have the knack of quickly homing in on the correct hypothesis, "after two or three or at the very most a dozen guesses"[74].

So, the claim is that if our use of abduction is governed purely by chance, the remarkable success of scientific inquiry over the last few hundred years[75] seems inexplicable. One obvious objection to the argument that leads to this conclusion is that it appears to misrepresent the phenomenology of scientific inquiry. Peirce writes as if the scientist investigating some new phenomenon is liable to be overwhelmed by the sheer number of potentially explanatory factors; when it comes to selecting a working hypothesis, she is spoilt for choice, and a kind of intellectual paralysis ensues. But typically, scientists struggle to come up with a single explanatory hypothesis when something unexpected shows up in their experiments. The reason for this is that the vast majority of factors that, in the abstract, *could* be relevant, such as the conjunction of the planets, or the day of the week, are simply ignored. Explicit attempts to justify the neglect of so many logical possibilities are seldom made, but the idea must be that neither our well-established general views about how the world works, nor existing scientific theories admit the possibility of astrological influence, and so forth. At any given stage of inquiry, the best available account of how the world works severely constrains the factors that scientists are inclined to incorporate into their explanatory hypotheses. Hence, feeling hemmed in, rather than limitlessly free, is characteristic of the search for scientific explanation.

This objection is telling, but not decisive. For presumably scientists have had to *learn* to restrict their attention to the possible explanatory factors that are close at hand:

[73] C. S. Peirce, *Collected Papers*, 5.172; section 5.591 contains remarks in a similar vein.
[74] C. S. Peirce, *Collected Papers*, 5.172.
[75] That science has made remarkable progress is something that, according to Peirce, no sane person can doubt; see *Collected Papers*, 5.172.

> How have they learned this? By an induction. Very well, that induction must have been based upon a theory which the induction verified. How was it that man was ever led to entertain that true theory[76]?

In other words, supposing that, at some point, scientists have recognised that they seldom have to look beyond local influences in order to account for some phenomenon merely defers the problem. This point is easy to miss because there is an almost irresistible tendency only to consider current scientific practice. But the assumption that such bizarre factors as Peirce mentions are rarely, if ever, explanatorily relevant is already deeply ingrained at this stage of the scientific project. As an antidote to this parochial line of thought, Peirce urges us to imagine the period before this hard-won insight occurred to anyone. He invites us to consider the "twenty or thirty thousand year period during which man has been a thinking animal", and asks whether the current level of scientific knowledge is explicable on the assumption that hypotheses for experimental testing are selected randomly. As Peirce has already eliminated the possibility that the logic of scientific inquiry compels us to take certain hypotheses seriously, there is only one conclusion to be drawn:

> ... man has a certain Insight, not strong enough to be oftener right than wrong, but strong enough not to be overwhelmingly more often wrong than right, into the... general elements of nature. This faculty is at the same time of the general nature of Instinct...[77]

As Peirce goes on to make clear, the faculty he alludes to in the above passage resembles instinct in two respects. Firstly, it jumps to conclusions that are in no way justified by reason, or by the sensory evidence available to us at the time. And secondly, it has a certain liability to error, even though the relative frequency with which it is right is, all things considered, "the most wonderful thing in our constitution".

There are time when Peirce gestures towards some kind of evolutionary account of how our intellectual faculties are attuned to the natural order around us. It is difficult to know how seriously to take this suggestion, though, since no sooner is it raised than Peirce sets it aside:

> You may say that evolution accounts for the thing. I don't doubt it is evolution. But as for explaining evolution by chance there has not been time enough[78]

[76] C. S. Peirce, *Collected Papers*, 5.591.

[77] C. S. Peirce, *Collected Papers*, 5.173.

[78] C. S. Peirce, *Collected Papers*, 5.172. Although Peirce seems to accept an evolutionary explanation here, many writers would regard the suggestion that evolution leaves room for a *non-random* element

It is hard to know what to make of this passage since Peirce appears to take back with one hand, what he has just given with the other. My feeling is that an evolutionary story *may* plausibly account for our grasp of what is nowadays known as "folk physics", but that the gulf between such primitive rules of thumb and the arcane, but sophisticated, conceptual framework of modern physics is simply too vast to be bridged. Peirce himself recognises this problem and seems ready to concede the point in one paragraph:

> As we advance further and further into science, the aid that we can derive from the natural light of reason becomes no doubt less, and less[79]

This claim, plausible enough when Peirce made it, can only be more so in the light of the major theories of twentieth century physics: general relativity, and, especially, quantum mechanics.

But now it looks like we have a problem again. An instinct for the truth was supposed to account for our ability to home in on the correct explanatory hypothesis, and this in turn was to be given an evolutionary explanation. But if the relevance of evolutionary considerations diminishes as science progresses, as it seeks to answer questions that are ever more distant from the primitive mode of life in which "folk physics" developed, what role is left for instinct? Fortunately, there is another side to Peirce's writings on abduction that I have not yet mentioned; according to some passages, the problem of abduction is best resolved, if that is the right word, by attending to the role that *regulative hopes* play in our inquiries.

The best way to approach this issue is to ask what the very activity of inquiry, of the sort that scientists engage in, must presuppose, if it is to make sense. I think that most people would accept that scientific inquiry involves some attempt to discern *order* in the range of natural phenomena we encounter in experience[80]. Is the existence of such order something that science itself demonstrates? Yes, insofar as we can point to past inquiries that have established the existence of natural order in a given domain. But no, insofar as scientific inquiry always involves a step into the unknown, an attempt to account for some fresh anomaly that does not quite fit within the existing theoretical framework. In such circumstances, we do not *know* that diligent inquiry will bring to light an underlying order,

as counter to the spirit of the theory. This point is made forcefully, and at length, in Daniel Dennett's *Darwin's Dangerous Idea*.

[79] C. S. Peirce, *Collected Papers*, 7.220.

[80] As we shall see, the idea, advocated in this chapter, that the goal of scientific inquiry is understanding accords with this intuition.

but it seems entirely reasonable to *hope* that it will. Indeed, inquiring blindly without, at some level, entertaining something like this hope, would be quite irrational.

Peirce's remarks on abduction may be seen as a description of the other side of the coin. In order for scientific inquiry to make practical sense, it is not enough to posit an orderly universe; it must also be supposed that scientists are capable of discovering that order, that their intellects are somehow attuned to it. In other words, as Peirce points out, the hypotheses that scientists are naturally inclined to find plausible, possible ways of ordering the phenomena, must stand a fair chance of being correct, otherwise scientific progress would grind to a halt. And again, the attunement of our intellectual faculties to the order that we hope is out there is something that we can only hope for[81]. In certain passages, Peirce seems to recognise that this is the case:

> But the saving truth is that there is a Thirdness[82] in experience, an element of Reasonableness to which we can train our own reason to conform more and more. If this were not the case, there could be no such thing as logical goodness or badness; and therefore we need not wait until it is proved that there is a reason operative in experience to which our own can approximate. We should at once hope that it is so, since in that hope lies the only possibility of knowledge[83].

In this passage, we see Peirce arriving at a very similar conclusion to Socrates about the role that hope plays in our inquiries, although via somewhat different routes. Socrates evoked the Theory of Recollection to defuse Meno's paradox, whilst conceding that we can only hope that the theory is correct. Peirce worries about the finitude of our intellectual powers compared with the vastness of the universe upon which we are called to exercise them, and sees some form of attunement between mind and nature as the answer, whilst recognising, in his more lucid moments that this is something that we can only hope for.

[81] Occasionally Peirce attempts to ground this hope in a naturalistic story. The basic idea is that since our minds are themselves a product of the operation of natural laws, there is an affinity between mind and nature that ought to enable the former to know the latter. This claim sounds pretty implausible, and, as I hope to show in the next chapter, is, in any case, not needed.

[82] One of Peirce's infamous categories, the other two being, not surprisingly, Firstness and Secondness. Interested readers should consult Peirce's, "On a New List of Categories" in N. Houser and C. Kloesel (eds.), *The Essential Peirce*, and C. Hookway, *Peirce*, Ch. 3.

1.4 Summary

The discussion of rational inquiry in this chapter has highlighted the following key points:

- Understanding is the goal of inquiry, or at least of many of the inquiries that we hold in the highest regard.
- Understanding is, at least partly, a result of trying to make sense of new information by seeing how it squares with prior beliefs and intuitions.
- When one has understanding of a given domain, one's beliefs enjoy a distinctive and valuable form of stability.
- The possibility of rational inquiry depends upon our entertaining hopes in things that are not amenable to rational demonstration.

The first and second points will be taken up again in chapter 3, and the third in chapters 4 and 5. Before all that, chapter 2 examines the theme of hope, and its place in rational inquiry in more detail.

[83] C. S. Peirce, *Collected Papers*, 5.160.

Chapter 2: Hope and Inquiry

2.1 Introduction

At the end of chapter 1, we saw that, for Peirce, the possibility of scientific inquiry depends ultimately upon the hope that nature has an intelligible order to which our abductive ability is attuned. Our earlier examination of the *Meno* suggested that the Theory of Recollection must also rest on similar assumptions, which, again, can be no more than hopes. This chapter develops the idea that such hopes play an important role in rational inquiry by examining the nature and legitimacy of hope in general.

Section 2.2 uses material from Aquinas to frame a conceptual analysis of hope, one that is subsequently applied to the fundamental hopes that underwrite the possibility of rational inquiry. Section 2.3 addresses the question of the legitimacy of hope. The conceptual analysis of the previous section provides a useful starting point for thinking about this issue, but the somewhat unusual hopes that play a part in rational inquiry require special treatment, one that relies heavily on the ideas of William James, as set out in his defence of "The Will to Believe". Section 2.4 uses some of Kant's less well known writings to give a more detailed account of the nature of these fundamental hopes and their legitimacy.

Much of the material in sections 2.3 and 2.4 amplifies and reinforces the conclusions of Plato and Peirce described in the previous chapter. But throughout these sections I also return to one of the outstanding issues from that discussion: is it advisable to base major commitments, such as the quest for scientific knowledge, on nothing more than a hope? In doing so, are we not simply guilty of misguided optimism and wishful thinking[1]? In reply, I suggest that although commitment to rational inquiry requires an attitude of optimism, it is not misguided, since the alternative is a form of epistemic despair.

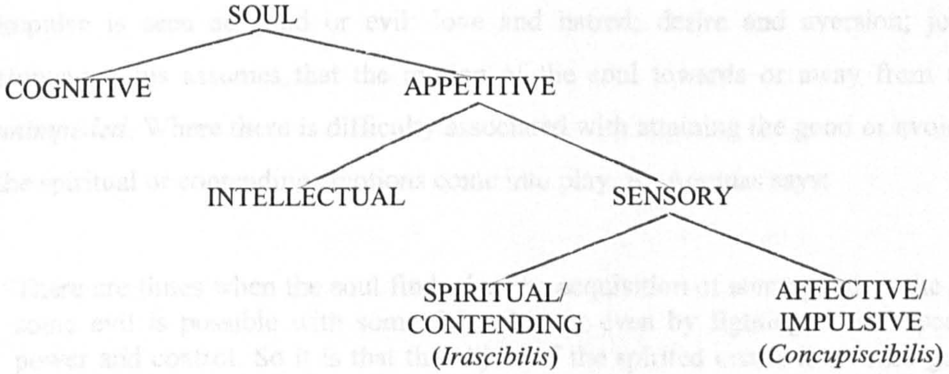
2.2 Towards a Definition of Hope

The main purpose of this section is to arrive at a working definition of hope that will: (a) apply to the special case of the hopes presupposed by rational inquiry; and (b) offer some preliminary clues about how the legitimacy of these hopes might be evaluated. As Aquinas presents a well worked out account of the nature of hope in the *Summa Theologiae*, I make

[1] Compare the attitude expressed by Samuel Johnson in describing the proposed second marriage of a friend as "the triumph of hope over experience". G. K. Chesterton articulated another variation on this

extensive use of his ideas. Before looking at these, it will be useful to have a rough idea of his classification of the emotions in general:

Figure 2.1: Aquinas' Taxonomy of the Soul



The main division within the soul is between the cognitive powers, and appetite, otherwise known as desire. Appetite is sub-divided into an intellectual and a sensory component. The sensory aspect of appetite concerns the immediate impact of objects upon our senses, and the various feelings they give rise to. The intellectual component of the appetite is engaged once these feelings are aroused, in what is essentially practical deliberation. Finally, Aquinas distinguishes within the sensory element of the appetite between affective or impulsive emotions and spiritual or contending emotions. As the name suggests, the impulsive emotions are modelled upon the movement of one body towards another under the influence of an attractive force, or impulse. Take the case of a particular good, to which we feel attracted. Our movement towards this good has three logically distinct aspects:

1. *Inclination* towards the good	Love	*Amor*
2. *Motion* towards attaining the good	Desire	*Desiderium*
3. *Repose* in the possession of the good	Joy	*Gaudium*

Likewise, in the case of a perceived evil:

popular theme when he defined hope as "the power of being cheerful in circumstances which we know to be desperate". Both witticisms are quoted in S. Sutherland's article, "Hope".

1. The evil produces a repulsion	Hatred	*Odium*
2. Motion *away* from the evil	Aversion	*Fuga*
3. The evil overtakes us	Grief	*Dolor*

So, the impulsive emotions come in three pairs, according to whether the cause of the impulse is seen as good or evil: love and hatred; desire and aversion; joy and grief. However, this assumes that the motion of the soul towards or away from the object is *unimpeded*. Where there is difficulty associated with attaining the good or avoiding the evil, the spiritual or contending emotions come into play. As Aquinas says:

> There are times when the soul finds that the acquisition of some good or the avoidance of some evil is possible with some difficulty, or even by fighting; it is beyond our ready power and control. So it is that the object of the spirited orexis is a sense-good or sense-evil *qua arduous*; i.e. in so far as its acquisition or avoidance involves some kind of difficulty or struggle[2].

Hence, these emotions are *contending*, or, in Latin, *irascibilis*, from which our word "irascible" is derived[3]. Aquinas identifies five contending emotions altogether: hope and despair, the objects of which are good; fear and courage, the objects of which are evil; and anger, which has no contrary. The rationale for this list is easily appreciated, given Aquinas' general conception of the contending emotions. Take the hope/despair pair. The object of both emotions is some particular sensory good which is difficult to attain. The emotions relate to the two aspects of this object: *qua* good, it attracts and arouses the emotion of hope; *qua* difficult to attain, it repels and occasions despair. Similarly with fear and courage: *qua* evil, the object is a source of fear, yet as the evil is difficult to avoid, it may also be the occasion of courage. Anger, the odd one out, is thought to be aroused when a courageous effort to avoid a recalcitrant evil fails.

Against this background, Aquinas' thoughts on hope can be stated relatively briefly. He says that hope is "a movement of appetite aroused by the perception of what is agreeable, future, arduous, yet possible of attainment"[4]. This definition identifies four features of the object of hope, each of which is meant to allow us to contrast it with one of the other emotions:

[2] Aquinas, *Summa Theologiae* 1a 2ae 23,1.

[3] Aquinas' language here is reminiscent of Peirce's talk of *roadblocks* on the path of inquiry, and the *irritation* which doubt creates in the mind of the inquirer. See Peirce, "The Fixation of Belief", especially sections III and IV.

[4] *Ibid.* Ia 2ae 40, 2.

1. It is *good*	Contrast fear
2. It is in the *future*	Contrast joy
3. It is *attainable*	Contrast despair
4. It is *difficult* to attain	Contrast desire

This is a good definition, but suffers from one general defect: it is expressed in terms of the *object* of hope rather than the *subject* who entertains the hope. But there are at least three reasons for supposing that hope is, in the language of contemporary philosophy, a *propositional attitude*. Firstly, the canonical form of an expression of hope is "S hopes *that* p", where S denotes an individual, or group of individuals, and p is a proposition which describes the object of hope[5]. Secondly, the above construction exhibits the kind of scope ambiguity that is characteristic of propositional attitudes. To see this, consider the sentence, "I hope that someone phones me tonight". Depending upon the scope of "hope that" and "someone", this could mean either:

There is a specific person whom I hope will phone	$(\exists x)$(I hope that x phones me tonight)
I hope that somebody (anybody!) will phone	I hope that $[(\exists x)$ x phones me tonight$]$

Thirdly, "S hopes that p" exhibits another characteristic feature of propositional attitudes, referential opacity. As long as I am blissfully ignorant, I can hope that my friend Ghengis visits me tonight, without hoping that the Broomhall Axe Murderer comes for tea, even though Ghenghis *is* the Broomhall Axe Murderer.

Given that hope is a propositional attitude, we need to replace each element of Aquinas' definition with one that involves an attitude of the subject. Doing so acknowledges that I may hope for something as long as I *believe* it to be good, in the future, and attainable (with difficulty), even though I may be mistaken on all counts. One consequence of this is that hope is no longer restricted to things which lie in the future. As long as I am suitably ignorant, I can hope for things which are present ("I hope that Ghenghis is home by now"), and past ("I hope that Ivan passed his driving test"). However, I think that Aquinas is right to believe that typically, hopes are future-orientated, and that this is something to do with the way the future is normally regarded as *uncertain*. Putting all of this together:

[5] Stuart Sutherland however raises some interesting questions about the primacy of the "S hopes that p" construction, especially where the hope is of a religious character. See his article "Hope". I believe that my conclusions, in this chapter, about the nature and legitimacy of the special hopes that underlie the possibility of rational inquiry are in agreement with the spirit of Sutherland's paper.

S hopes that p iff:

1. S desires that p *or* S believes that p is good.
2. S believes that p is uncertain.
3. S believes that p is possible *or* has no reason to believe that p is impossible.
4. S believes that p is difficult to attain.

The first three conditions on the list are fairly straightforward. Firstly, people do not usually hope that evil befalls them. They may hope that bad things happen to someone else, but this is because they believe that another's misfortune will make them happy. A complication arises when p, which S would not normally desire in itself, is a necessary condition for q, which S desires. In such cases, it may nevertheless be true that S hopes that p. For example, if my computer is on the blink, I may find myself hoping that the house is burgled, so that I can claim on the insurance and buy a new PC, even though I would not normally want such a thing to happen. I think that the best way of dealing with these cases is to insist that, in them, S really does desire that p. In the example, I do want my house to be burgled, this seemingly perverse desire being understandable in the circumstances. An alternative response to these cases is to point out that there is still some connection between hope and desire, even though it is indirect.

The second element of the definition is less contentious. It is intended as a weaker version of Aquinas' second condition, since, once we adopt the idiom of propositional attitudes, it is clear that something does not have to be in the future in order to be uncertain. The essential thought is that when p is regarded as certain, adopting an attitude of hope towards it is inappropriate on the grounds of redundancy.

The third condition is really the converse of the second: just as hope is redundant when p is regarded as certain, so is hope futile when p is thought to be impossible. As Aristotle sagely comments, "When men come upon something impossible, they turn away"[6]. This feature distinguishes hope from wishful thinking, for even if I am convinced that p is impossible, I can still wish that things were different. It is also worth noting that, in this instance, "possibility" covers both the logical and the physical varieties: it is as futile for me to entertain hopes of being able to levitate one day, as it is for me to hope to square the circle.

[6] *Ethics* III, 3, 112, b24.

The final condition is the most contentious, but also the least essential as far as I am concerned. It seems plausible because there appears to be something inappropriate about saying, for example, "I hope that I can lift this cup of rosehip tea", when there is nothing to stop me. However, I do not see that there is any *logical* contradiction here: as long as the object of desire is not absolutely certain, then hope is a possible attitude to adopt. However, the tentativeness I express about my ability to lift the cup may be criticised on two grounds. Firstly, it seems irrational, or at least highly neurotic, to worry about what may happen to the extent that even lifting a cup of tea comes to seem problematic. And secondly, it may lead my audience to believe - incorrectly - that there is something that might prevent me from lifting the cup. That is, although there is nothing logically wrong with my statement, it is likely to generate misleading conversational implicatures[7].

The fourth condition also suffers from another defect: the word "attain" suggests that p describes a state of affairs that could be brought about, either by my own efforts or by somebody else's. Although this is appropriate in cases like "I hope that I pass my driving test", or "I hope that England beat Australia", it is not generally the case. Having bought a lottery ticket, I can certainly hope that my numbers come up even though there is nothing (this side of the law) that I, or anyone else, can do to bring about this happy state of affairs. This kind of case indicates that we would be better off with some kind of low probability requirement as our fourth condition. However, this suggestion raises problems of the "How low is 'low'?" variety that would, I believe, be fruitless to pursue.

For both of the above reasons, I prefer to simply leave out the fourth condition, and work with the following definition:

S hopes that p iff:

1. S desires that p *or* S believes that p is good.
2. S believes that p is uncertain.
3. S believes that p is possible *or* has no reason to believe that p is impossible.

Whilst noting that p will often, if not usually, describe a state of affairs that is difficult to bring about, or that is unlikely to be the case for other reasons.

[7] See Grice "Logic and Conversation", for more about conversational implicature.

Before considering an important objection to the above analysis, let's pause to confirm that it applies to the fundamental hopes we identified in chapter 1. Suppose that S is a scientist inquiring into some aspect of the physical universe. According to Peirce, S must, at some level, entertain the idea that nature is orderly, and that our minds are attuned to that order, even though this is something that she can only hope is the case. Is this an appropriate object of hope according to our definition? It looks like it is, for:

1. From S' point of view, it is certainly desirable that nature is orderly and that we are attuned to that order, otherwise further inquiry is likely to be futile.

2. Unless S knows something that we don't, she cannot claim with certainty that the universe is orderly or that we have some special insight into the nature of that order.

3. Nevertheless, S has no reason to suppose that it is impossible that nature should be orderly. Likewise, as far as S is concerned, there is no reason why our minds should not, with sufficient effort, be able to latch on to that order.

As to whether there is some difficulty that would hinder or prevent the existence of order in the natural world (the fourth condition of our original definition), it seems impossible to say.

It also looks as if the Theory of Recollection[8] that Socrates uses to motivate an inquiry into the nature of virtue is an appropriate object of hope:

1. According to Socrates, it is highly desirable that the Theory of Recollection be true (or at least that we believe it to be true), since the alternative is a kind of intellectual apathy and defeatism.

2. As we saw, Socrates is far from certain that the Theory of recollection is true, having only heard of it from other, albeit wise, people.

3. Nevertheless, Socrates has no reason to suppose that the Theory of Recollection could not possibly be true.

So, as well as being inherently plausible, our definition of hope ties in well with the earlier discussion of fundamental hopes and their role in inquiry.

However, there is at least one significant problem with the above analysis of hope. Exploring this difficulty will reveal further connections with the material in chapter 1, and lead in to the question of when hope is legitimate. The problem is that the three conditions in our definition are not jointly sufficient, for they are entirely compatible with the presence of despair, the antithesis of hope[9]. To see this, consider the following example. Suppose that Genghis falls ill, and is rushed to hospital for a life saving operation. Doctors inform his parents that the operation is by no means certain to succeed, but it is their son's only chance. Clearly both parents want Genghis to pull through, and both are aware that it is possible, but not certain that he will do so. In this situation, all of the above criteria for hope are satisfied. The problem is that, on hearing the prognosis, Genghis' mother adopts an attitude of resolute hope, whereas his father falls into a state of abject despair. Nothing we have said so far excludes this possibility.

This suggests that we need to add another condition to our list. One possibility is to insist that the states of hope and despair just *feel* different; but this does not sound very explanatory. Another idea is to appeal to behavioural criteria: hope and despair differ in the way that they manifest themselves in behaviour, or, at least, in dispositions to behave in certain ways. To return to our example, Genghis' mother's hopeful attitude may be shown by her willingness to make plans for a future in which her son is present; she redecorates his bedroom and the like. On the other hand, the despair which the father feels will probably prevent him from engaging in such activity. The problem with this suggestion is that it is possible to imagine the father going along with everything the mother says and does, but in a half-hearted manner. His mood is one of despair, even as he hangs the new wallpaper in his son's room.

Although there may be some truth in both of the above suggestions, I think we can say something more illuminating about the distinction between hope and despair. *Ex hypothesi*,

[8] From now on, I shall use "Theory of Recollection" as shorthand for both the Theory of Recollection proper (the idea that the soul has already learned all truths in a previous existence, but forgotten many of them), and the associated claim that "all of nature is akin".

[9] Note that the objection to be discussed below does not show that any of the conditions in our definition of hope are not necessary; it merely shows that they are not jointly sufficient. And the fact that I do not propose an additional condition at the end of the following discussion, does not make the existing conditions any less necessary.

in the above example, Genghis' mother and father have a common desire, and have the same beliefs about the probability that this desire will be satisfied. Where they differ is in the way they *regard* the evidence concerning the likely outcome of the operation. His mother is somehow able to dwell on the possibility of a favourable outcome, whereas his father cannot stop himself from dwelling on the crisis' tragic possibilities. The underlying idea here is none other than that behind the old saying that the optimist sees a half-full glass whereas the pessimist sees the same glass as half-empty. The existentialist catholic philosopher Gabriel Marcel makes the point rather more poetically when he remarks that "to trace the visage of despair is to conceive the features of hope"[10]. Wherever there is uncertainty, this possibility of imaginatively anticipating the future course of events arises. Where the future possibilities are regarded not with neutrality, but with longing or aversion, there also arises the possibility of hope or despair[11, 12].

What does this tell us about the nature of hope? Remember that in our example, both parties agree on the objective features of the crisis - their son is ill, recovery is possible but not certain - just as the optimist and pessimist both agree that the glass contains so many millilitres of liquid. This suggests that, as long as certain minimal conditions are fulfilled[13], there is only a loose correlation between what an individual believes to be the case and her adoption of an attitude of hope rather than despair - and that she cannot easily be accused of being irrational for adopting one attitude rather than the other[14]. To return to our example, both of Genghis' parents appear equally rational in terms of the beliefs they form on the basis of the evidence available to them. Genghis' dad's despair cannot be dismissed as irrational since there is a significant chance that Genghis will not survive; but, by the same token, Genghis' mum is just as rational[15], given that the operation may succeed.

[10] *Homo Viator*, p105. The point also brings to mind Wittgenstein's work on aspect perception in *Philosophical Investigations*, II xi. From this perspective, his mother may be said to *foreground* the evidence that suggests Genghis will recover, whilst his father allows this evidence to recede into the background.

[11] It may be said that this is just a special case of the idea that we distinguish between hope and despair on behavioural criteria - if thinking about something may be described as a form of behaviour. Fine, but it is still a *special* case, one worth emphasising.

[12] Perhaps it is not surprising that we cannot reduce hope and despair to combinations of beliefs and desires, and use this analysis to distinguish between them. Both, after all, are usually classified as emotions, and it is at least plausible to suppose that emotions, in general, cannot be reduced to combinations of beliefs and desires.

[13] Namely those set out in our analysis of hope.

[14] Here, we begin to move on to the question of when hopes may reasonably be entertained. This will be discussed in much greater detail in section 2.3.

In the above example, the adoption of a particular attitude - hope or despair - is determined not so much by the evidence, as by something like individual temperament. Genghis' mum, we might say, just happens to have a naturally "sunny disposition", whereas his dad is apt to look on the black side; and this difference accounts for their contrasting reactions to the bad news about their son. Since people do not have a lot of control over their temperament[16], it would be misleading to say that one parent chooses to remain optimistic whilst the other opts for despair. However, in some circumstances adopting an attitude of hope rather than despair might be something that one could choose to do. Socrates seems to imply that this possibility exists when he commends the Theory of Recollection to Meno on the grounds that accepting it will encourage him to a life of rational inquiry, thereby making him a "better, braver, and more active"[17] man. In the previous chapter, we noticed that there was something odd about the Theory of Recollection, and the way it is introduced into the *Meno* dialogue, eventually tracing this to the fact that the theory is offered as an object of hope rather than belief. To this we can now add that the theory is something that we can *choose* to place our hope in, if the activity of rational inquiry is of sufficient importance to us[18].

In examining the distinction between hope and despair, we have found ourselves proposing answers to the question of when hope is a legitimate attitude to adopt. The next section addresses this question directly, and expands upon the tentative answers given so far.

2.3 Legitimate Hopes

2.3.1 Aquinas on Hope and Experience

For a cynical view of the legitimacy of hope, we need only recall Johnson's witticism about a second marriage being the triumph of hope over experience. The implication is that hope and experience are entirely independent of, and even antagonistic to, one another. Aquinas, as might be expected, takes a less cynical line. Although his ideas about the legitimacy of hope have limited application to the fundamental hopes that play a part in rational inquiry, it is still worth briefly reviewing his conclusions for the jumping off point they provide. Let's begin by taking another look at the definition we extracted from his work[19]:

[15] If not more so, if we broaden our conception of what constitutes rationality. See section 2.3 for more details.

[16] At least in the short term; it may be possible to cultivate an optimistic outlook over a period of time.

[17] *Meno*, 86b.

[18] Temperament will have a role to play as well, but there certainly appears to be more scope for choice here than in the Genghis example.

S hopes that p iff:

1. S desires that p *or* S believes that p is good.
2. S believes that p is uncertain.
3. S believes that p is possible *or* has no reason to believe that p is impossible.

Notice that, within this definition, it is possible to distinguish between an affective component, the desire that p, and a cognitive component, the beliefs about the likelihood of p. When Aquinas comes to examine the legitimacy of hope in the section of the *Summa* that deals with the relationship between hope and experience[20], he concentrates, quite naturally, on the cognitive aspect of the above definition. Experience modulates what we may hope for in that it has an effect upon what we *believe* to be possible or impossible[21]. Bearing this in mind, Aquinas suggests that there are three ways in which experience can influence the reasonableness of our hopes[22]:

- Experience may suggest that a particular good is attainable; this permits and enables us to entertain hopes of attaining that good in the future.

- Experience may show that what we previously hoped for is impossible; the rational policy in these circumstances is to abandon all of our hopes in that direction.

- *Lack of* experience may cause us to persist in hoping for a particular good when it is in fact impossible[23].

An example of each of these cases in turn:

- Although I have not won the National Lottery, I know that at least one person each week has done so, and am therefore justified in believing that success is possible. Hence, having bought a ticket, I am surely justified in hoping that my numbers come up.

[19] Of course, we saw at the end of section 2.2 that this definition is equally compatible with despair.
[20] *Summa Theologiae*, 1a 2ae 40, 6
[21] Or, perhaps more generally, experience can provide us with information about the probabilities and relative probabilities of certain events.
[22] Given the compatibility of our definition with despair, we would expect similar remarks to apply in that case; for example, experience may alleviate despair by suggesting that, contrary to what we thought, a particular good is attainable, after all.

- Now that a proof of Fermat's last theorem exists, it is no longer rational of me to hope to become famous by discovering a counterexample to it.

- If I don't know of Andrew Wiles' historic proof (I missed *The Guardian* that day), I may legitimately continue to entertain hopes of disproving Fermat's last theorem. Although I may be badly misinformed[24], I am not behaving irrationally - unless someone brings the proof to my attention and I carry on regardless.

These brief remarks of Aquinas about the relationship between hope and experience are fine, as far as they go; but, unfortunately, that is not very far. In particular they shed little light on the question of whether the kind of fundamental hopes discussed in the previous chapter may reasonably be entertained. There are, I believe, two reasons for this.

To appreciate the first reason, look again at the first of the three examples above. I have bought a lottery ticket, and given that I know there have been previous winners, I am justified in hoping for similar good fortune for myself. But notice what an etiolated specimen this hope is. By this I mean that, although it's reasonable for me to think "I hope I win the lottery" to myself, or say it to my friends, I would not be justified in, say, going out and buying everybody speedboats on the basis of that hope. It would be deeply irrational to take such extravagant and irrevocable action when the probability that what I hope for will materialise is so low. It is tempting to say that the hope that I shall win the lottery doesn't really amount to much because there is so little that I can*do* with it, or on the basis of it. It is reasonable, in some minimal sense, for me to entertain the hope, but so what?

This overstates the case a little. Although, the hope that I shall win the lottery isn't good for very much, it may be good for something. For example, it might be reasonable of me to *promise* my best friend a speed boat *if* I win the lottery. Making such a promise is probably acceptable, given the nature of the hope in this particular case. At the other end of the spectrum, it may be quite reasonable to take decisive action if it is likely that the object of our hope will materialise. For example, it is probable that a good friend, who I do not get to see very often, will be in town this weekend. Naturally, I hope that we shall meet up. But I am also justified in acting on the basis of that hope - keeping the weekend free, buying a

[23] Hence as Aquinas perceptively points out, *hope abounds in the young and the inebriated* as a result of a paucity of, and a dislocation from, experience, respectively.

[24] Which, in the circumstances, is bad enough, of course.

good bottle of wine, and so forth. These examples suggest that the degree to which it is rational to *act* on the basis of a hope (as opposed to merely entertaining the hope in an abstract fashion) depends upon:

- The likelihood that what is hoped for will turn out to be the case; the more likely this is, the more rational it is to act on the strength of the hope.

- The nature of the action being contemplated. Other things being equal, it is less reasonable to take irrevocable or costly actions on the basis of a hope than it is to take actions that are modest and easily taken back.

The main point to emerge from the above discussion is that often we are interested not so much in the rationality of entertaining a hope, which, as the lottery example shows, can be a relatively trivial matter, as in the rationality of *acting* on the basis of that hope. This is certainly the case as far as the fundamental hopes that underlie the possibility of rational inquiry are concerned. Peirce's scientific inquirer, and Socrates' philosopher are both contemplating a course of action that involves a serious commitment in terms of time and effort from them. Are they justified in making such a commitment on the strength of nothing more than a hope that nature is orderly, or that the Theory of Recollection is true? This, I believe, is the important question; but it is not one that Aquinas' remarks about the relationship between hope and experience will answer.

There is a second reason why Aquinas' ideas about the legitimacy of hope are of limited use to us. It is clear that his remarks only apply to those cases where experience *could* tell us something about whether the object of our hope is possible or impossible[25]. We might call such hopes *mundane* to reflect the fact that evidence regarding the possibility or otherwise of their objects might be encountered within the field of possible experience[26]. The hope that I shall win the lottery, though somewhat extravagant, is mundane in this sense, as there is

[25] Later, Aquinas discusses hope *qua* theological virtue; that is, as it bears upon our relationship with God. Arguably experience has little, if anything, to tell us about whether the existence of such a transcendent being is genuinely possible. Such a hope therefore counts as non-mundane, according to my definition. Unfortunately, Aquinas says nothing about how he believes we should evaluate this kind of hope.

[26] To use some premonitory Kantian terminology.

evidence that justifies my belief in the possibility of winning, namely other people's good fortune[27].

The problem is that the hopes that underlie the possibility of rational inquiry do not seem to be mundane in this sense. For Peirce, the hope in question concerns the nature of the universe as a whole, and the intellectual powers we exercise in seeking to understand it; and for Socrates it is hope in the Theory of Recollection. In neither case is it clear what *experience* can tell us about the possibility or impossibility of the object of our hope. The best that we can say[28] is that there is no evidence to suggest that the objects of these fundamental hopes are impossible. But then the question of whether this is really enough to base such an important commitment on just resurfaces.

The remainder of this chapter addresses these two shortcomings (from our point of view) of Aquinas' account of the legitimacy of hope. The material on William James and "The Will to Believe" in section 2.3.2 is mainly concerned with the relationship between hope and action. Section 2.4, which explores Kant's account of hope, has a lot to say about the non-mundane character of the hopes that interest us, whilst also amplifying some of the other themes of this chapter.

2.3.2 William James on "The Will to Believe"

Despite the title, "The Will to Believe", William James' celebrated defence of the right of an individual to religious faith, contains insights that shed light on the fundamental hopes that concern us, and on their place in our lives. In particular, a lot of what James says has implications for when we might be entitled to act on the basis of such hopes[29].

The "Will to Believe" was originally presented as a talk to the Philosophical Clubs of Yale and Brown Universities. James believes that a typical member of this highly educated audience will have a narrow scientific outlook, and will be unsympathetic to religious belief as a result. If challenged, such an individual could try to defend this sceptical attitude by appealing to what might be called the "evidentialist" thesis:

[27] Though I suppose there could be a question about whether this is evidence for the claim that it is possible that *I* shall win the lottery (maybe I'm just fated not to).

[28] And did say when we applied our analysis of hope to these special cases near the end of section 2.2.

[29] I shall consider whether James should be talking about hopes rather than beliefs towards the end of this section. Kant also has some interesting things to say about this matter; see section 2.4 for more details.

> *We ought to give to every proposition which we consider as nearly as possible that degree of credence which is warranted by the probability it acquires from the evidence known to us*[30].

This expresses the intuitively appealing thought that a belief is rational to the extent that it is justified by the available evidence[31]. James notes that, in the hands of some of its advocates, the expression of this idea assumes a quasi-ethical, and somewhat shrill aspect. For example, in "The Ethics of Belief", W. K. Clifford claims that "it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence". The rigidity of this attitude bothers James. The "The Will to Believe" may be seen as his attempt to set out a more nuanced account of when a belief may legitimately be entertained, one which gives breathing space to non-evidentialist considerations. James accepts that the evidentialist policy is adequate in most circumstances, but insists that its adoption ought not to be enforced in all cases. It turns out that religious faith is one of these exceptions.

In order to understand why James thinks this, it is necessary to be familiar with some of his terminology. I summarise this below:

Hypothesis: James defines a hypothesis as "anything that may be proposed to our belief"[32]. I take it that this covers all meaningful sentences, but it could also include some which, upon further inquiry, prove to be meaningless.

Live Hypothesis: A live hypothesis is "one which appeals as a real possibility to him to whom it is proposed"[33].

Dead Hypothesis: The opposite of "live", obviously. A dead hypothesis is one which does not appeal to us as a real possibility.

Although James is not very clear about what the distinction between a live and dead hypothesis amounts to, it is possible to say a little about what he has in mind. His example of a dead hypothesis is "believe in the Mahdi", which, he assumes, does not appeal to anyone

[30]Russell, "Pragmatism", quoted by E. K. Suckiel in *The Pragmatic Philosophy of William James*, p73. The term "evidentialism" is taken from her discussion.

[31] In section 2.4, we shall encounter Kant's image of the scales of reason: evidence on one side, belief on the other. The idea he considers - that all bias should be eliminated from the scales of reason - is another expression of the evidentialist thesis.

[32] William James, "The Will to Believe", in G. Bird (ed.), *William James: Selected Writings*, p250. All page references will be to this volume.

in his audience as a real possibility. A corresponding live hypothesis would be "believe in Christ". As James acknowledges, this means that the vitality of a hypothesis is not an intrinsic property, but depends on the audience: an Arab might take the idea of following the Mahdi seriously. James also says that that the vitality of a hypothesis is indicated by one's readiness to *act* upon the assumption that it is true. Willingness to act irrevocably indicates maximum vitality, and, practically speaking, this is the state of belief. Finally, James says something about how hypotheses come to be live or dead for us in the first place. He gives the name "authority" to "all those influences, born of the intellectual climate, that make hypotheses possible or impossible for us, alive or dead". This suggests that whether a hypothesis is live or dead for us is, at least partly, determined by our culture.

Option: A choice between two or more hypotheses.

Live Option: A live option is one in which all of the alternative hypotheses are live.

Forced Option: An option is forced if there is no way of avoiding choosing one of the competing hypotheses. For example "either accept the Special Theory of Relativity as true or reject it as false" is not a forced option because I can always suspend judgement. However, "either accept the theory, or do without it" is forced, as is any option where the alternatives form a complete logical disjunction.

Momentous Option: A momentous option is one which is unique, where the stakes are high, and where one's decision is irreversible. For example, a once in a lifetime chance to join Dr. Nansen's North Pole expedition is a momentous option. The opposite of "momentous" is "trivial"

Genuine Option: An option is genuine if and only if it is live, forced and momentous.

With this terminology in place, we are now in a position to state the conclusion which - *contra* the evidentialist thesis - James wants to establish:

[33] *Ibid.* p250.

> *Our passional nature*[34] *not only lawfully may, but must, decide an option between propositions, whenever it is a genuine option which cannot by its nature be settled on intellectual grounds*[35].

Note how cautious James' conclusion is. It is not, as some philosophers claim, a license to ditch all of our evidentialist scruples and indulge in wishful thinking[36]. Rather, James carefully restricts his conclusion to those cases where the evidentialist principle cannot help us because the option before us "cannot, by its nature, be settled on intellectual grounds". In addition, the option must be genuine: live, forced, and momentous. Whether or not pixies play hoopla at the bottom of the garden when no one is watching[37] may be intellectually undecidable, but, as an option, it is neither live, forced, nor momentous, and it would not be rational to seriously entertain the idea.

The main reason why our 'passional nature' is necessarily implicated when choosing between options of the kind that James describes is that:

> *... to say, under such circumstances, "Do not decide, but leave the question open", is itself a passional decision - just like deciding yes or no - and is attended with the same risk of losing the truth*[38].

To appreciate this, we need to bear in mind that certain risks accompany any inquiry: the risk that we shall end up believing something that is false; and the risk that we shall end up not believing something that is true. How we respond to these risks is, James says, itself an aspect of our "passional nature", which, for want of a better term, I shall call *epistemic temperament*. For the sake of simplicity, we may take it that the range of epistemic temperaments forms a continuum: at one end we find the inquirer who is motivated by the

[34] Throughout, James uses "passional nature" and "willing nature" more or less interchangeably. The essay is called "The *Will* to Believe", whereas its conclusion refers to the determination of belief by our "*passional* nature". Elsewhere James says that by "*willing* nature" he means "all such factors of belief as *fear* and *hope, prejudice* and *passion*, imitation and partisanship, the circumpressure of our caste and set". This list includes what most people regard as cultural, as well as emotional factors. Perhaps James wants to apply the label "passional" to everything that has not (yet) been subjected to critical scrutiny. These factors determine the live hypotheses available to us, and may even influence which of these we accept.

[35] *Ibid.* p256.

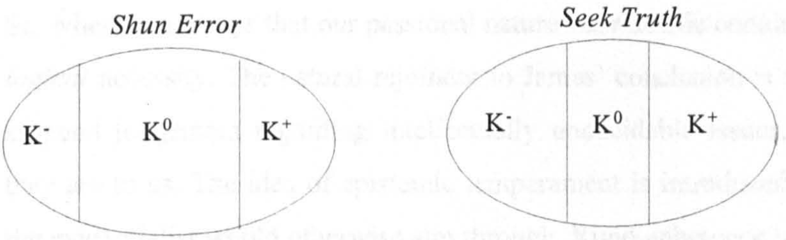
[36] Bernard Williams, who, in "Deciding to Believe", argues that one cannot regard something as a belief if there is so much as a suspicion that one acquired it at will, may be thought of as a relatively moderate member of this camp. James leaves himself open to such arguments through his tendency to conflate the will and the passions (see the footnote regarding our "passional nature" above). Despite the title of the essay, it is the latter rather than the former that James believes may, in special cases, legitimately influence our beliefs.

[37] Leaving no traces, of course.

[38] *Ibid.* p256.

desire to avoid error at all costs; whilst at the other we find the inquirer who is motivated by the desire not to let the truth get away, even if this means believing more falsehoods. To get a feel for what a difference epistemic temperament makes in practical terms, consider a simplified model of belief formation in which, given a proposition, p, I can only adopt one of three attitudes towards it: accept p as true; reject p as false; or suspend judgement regarding p. Then, the diagram below represents the two types of epistemic temperament just described:

Figure 2.2: Epistemic Temperaments



In the above diagram, K^- denotes propositions rejected as false, K^+ propositions accepted as true, and K^0 propositions regarding which the inquirer suspends judgement. The more conservative inquirer who shuns error suspends judgement on a wide range of issues where there is nothing, or very little to go on, or where the available evidence is inconclusive. The more speculative inquirer is prepared to run the risk of believing falsehoods for the sake of acquiring more truths, and therefore leaves fewer issues undecided.

When p is intellectually decidable, the temperament of individual inquirers does not matter in the long run. Although the evidence may be inconclusive now, further inquiry ought eventually to settle whether we should accept p or reject p. So, even if the conservative inquirer suspends judgement about p, sooner or later she will be able to assign p either toK^+ or to K^-. Likewise, if the speculative inquirer has already assigned p to either K^+ or K^-, further investigation will either vindicate this decision or reveal it to be a mistake. For intellectually decidable options, all inquirers, regardless of their temperaments, converge upon the same set of beliefs[39] in the long run[40].

[39] A set of beliefs that Peirce, for one, would identify with the true description of reality; see "How to Make Our Ideas Clear" in Houser and Kloesel (eds.) for his account of truth as the product of an idealised process of inquiry.

However, if p is intellectually undecidable, then, *ex hypothesi,* no amount of further evidence will enable the conservative inquirer to assign it to K^+, or K^-. By the same token, if the speculative inquirer takes p to be true, then no evidence can ever force him to revise his opinion. Hence, there is no convergence of opinion in the long run, and the risk of believing a falsehood or losing the truth forever is genuine. Question: is it rational to suspend judgement in these circumstances, and can the individual who refuses to suspend judgement be accused of irrationality? Answer: *in extremis* the issue of rationality does not seem to arise; adopting either policy is a decision to act in a certain way, and is more a reflection of one's temperament than of one's rationality (narrowly conceived).

So, when James says that our passional nature *must* decide certain issues, he is referring to a *logical* necessity. The natural rejoinder to James' conclusion is that it is always possible to suspend judgement regarding intellectually undecidable issues, no matter how important they are to us. The idea of epistemic temperament is introduced to close this loophole that the evidentialist would otherwise slip through. Rigid adherence to an evidentialist outlook is itself the mark of a conservative epistemic temperament, and thus betrays the influence of the passions. The strict enforcement of the evidentialist policy in all circumstances may be nothing more than intellectual *hubris*.

James expects his argument to encounter resistance, and attributes this to the grip the evidentialist thesis has on our imaginations. To loosen this grip, James describes a couple of cases where the influence of our passional nature upon belief formation appears to be rationally defensible. In the first of these, a chasm opens up on the mountain path I am walking on[41]. An avalanche prevents me from retracing my steps, and it is snowing heavily. No one knows where I am, and my food is running out. In short, I am lost, unless I can summon up the courage to attempt the leap across the chasm. Not far on the other side, I know, is a small village where I can recuperate and replenish supplies. The problem is that all of the evidence I possess implies that the jump is well beyond my physical capacities. Let

[40] Actually, this is probably not quite correct. Maybe some attitudes are just *too* speculative or conservative. The danger of being too speculative is that one would end up believing *many* false beliefs, and these could *destabilise* one's inquiries. This, it will be recalled, is one of the concerns that the Theory of Recollection is supposed to address (see pp15-19). The converse danger - being too conservative - is that one's inquiry will never get started (perhaps Descartes is in this situation as a result of applying the method of doubt). James' fundamental point is that there is a *zone of stability* within the set of all possible ETs, and as long as we remain within that zone, we cannot be accused of conducting our inquiries irrationally.

us say that I have never jumped more than six foot, and the chasm is eight foot across. According to James, in this grim scenario, the rational course of action is to make a last ditch effort and jump. Moreover, things are more likely to turn out well for me if I actually*believe* that, contrary to all of the evidence, I can succeed before I make the attempt. That is, I am more likely to succeed if my belief formation policy is not an evidentialist one on this particular occasion.

Another type of case favourable to James' thesis occurs often in the field of interpersonal relationships. Suppose that I have just met Genghis, and am wondering whether he really likes me. So far, there is hardly any evidence to go on, but what little there is does not settle the matter one way or the other. If I allow my conduct when I am with Genghis to be governed by the evidentialist principle, I am not likely to take any risks to get to know him. As a result, I will probably come across as rather detached and indifferent, and am likely to remain a stranger to him. The alternative is to have faith that Genghis likes me in advance of evidence that would support the belief, and to act upon this faith. If I allow my conduct when I am with Genghis to be guided by an attitude of trust, friendship may well blossom after all.

Both of the above examples strike me as persuasive, but it is worth making a couple of points about them before continuing. Firstly, they are both naturally taken as cases of faith *creating* the fact, rather than cases of faith making us *receptive* to facts we otherwise would not discover[42]. In this respect they appear dissimilar to the case of religious belief, where, it is traditionally assumed, faith makes us receptive to a level of reality which exists independently of, and prior to, ourselves. They are also, perhaps, dissimilar to the fundamental hopes[43] that underlie the possibility of inquiry, since these hopes posit an order that is likewise usually supposed to exist independently of, and prior to, our inquiries[44].

[41] Although this case is often discussed in connection with "The Will to Believe", it is in fact described in detail in "The Sentiment of Rationality". See G. Bird (ed.), *William James: Selected Writings*, pp20-53, and especially p43ff.

[42] Though the Genghis example could be read in the latter way, if we suppose that Genghis likes me from the outset. The distinction between faith making the fact, and faith making us responsive to the fact comes from E. Suckiel's discussion of "The Will to Believe", in *The Pragmatic Philosophy of William James*.

[43] Presently, I shall suggest that James' conclusion, originally phrased in terms of what we may *believe* in advance of the evidence, is best thought of as describing what we may legitimately hope for; or, at any rate, that a similar conclusion applies to the hope case.

[44] However, my thesis does not depend upon such a realist interpretation of these hopes. It may be, for example, that the vigorous pursuit of inquiry is somehow responsible for creating the order sought; and that the experience this engenders of living in an orderly universe makes for a better life than would otherwise obtain.

To appreciate the second point, we must recall that, according to James, *two* criteria must be satisfied before it is lawful and necessary for one's passional nature to influence belief formation: the option must be genuine *and* intellectually undecidable. The options described in our examples may reasonably be regarded as genuine, one being quite literally a matter of life and death, but are they really intellectually undecidable? That depends, I think, on whether we mean absolute undecidability, or undecidability for all practical purposes in the given context. As far as religious faith is concerned, the issue is almost certainly undecidable by rational means *in principle*; and the same could plausibly be said about the fundamental hopes that underlie the possibility of rational inquiry. However, the chasm jumping and interpersonal relationship cases seem much less resistant to intellectual inquiry. James' point in describing such examples must be that, although inquiry may be able to settle the issue *eventually*, a decision needs to be made *now*, in advance of the evidence that such an inquiry would turn up.

Although disanalogies between the examples James describes and the case of religious faith do exist, I do not think this undermines James' case. In the first place, his use of the concept of what I have called epistemic temperament already provides a general argument for the conclusion he wishes to establish. The specific examples he discusses are really just there to persuade the evidentialist within us that his conclusion is not as outlandish or dangerous as it might appear. Secondly, to the extent that his examples are disanalogous to the case of religious faith, the disanalogies are in James' favour: faith creates the facts in the examples, but merely makes us sensitive to them in the religious case; the options described in the examples are intellectually undecidable for practical purposes, whereas the religious hypothesis is undecidable in principle. In either case, we ought to be more, not less, inclined to accept that allowing one's passions to influence one's beliefs in the religious case is rationally defensible.

To summarise, James suggests that, *pace* the evidentialist, it is occasionally legitimate for our beliefs to be influenced by factors other than the balance of evidence. In particular, what James calls our passional nature has a part to play in determining the beliefs we may reasonably hold, as long as certain conditions apply: the option under consideration must be genuine and intellectually undecidable. He has a general argument for this conclusion that makes use of the concept of epistemic temperament, and backs up his conclusion with a number of examples that make it appear plausible.

James goes on to show that the case of religious faith satisfies the conditions he has described, and that, therefore, someone who adopts such a faith cannot be criticised as irrational. However, I want to argue that the fundamental hopes discussed in this and the previous chapter also satisfy James' criterion. Consider first the option between *believing*[45] that the natural world is orderly, and that our intellectual capacities are attuned to that order, and not believing the same. As presented, the option is forced since we must either adopt the belief or do without it. The option is also a live one, at least as far as most of the people reading this thesis are concerned. It certainly seems possible that the universe exhibits the kind of order Peirce describes, but the idea that we might some day have to reject this possibility is not outlandish. Thirdly, the option is momentous, on any reasonable interpretation of that word, being concerned with the nature of the universe and our place within it. The position we take on this issue will surely have a profound effect on the way we live the rest of our lives. So, the option described is live, forced and momentous, and therefore qualifies as a genuine option as far as James is concerned. As it also looks like a pretty good candidate for intellectual undecidability[46], it looks like the option is one we can legitimately leave to an individual's passional nature to settle. For example, someone with a passion for inquiry, and, perhaps, an optimistic temperament will happily accept the belief in an orderly universe, and live her life by it - and who are we to say that she is misguided?

Much the same can be said about the option between believing the Theory of Recollection, and not believing it. The option is forced, as before, and, we assume, it is live in the context in which it arises (though it may not be live for many people these days). It's also a momentous option, for the same reason as before, and is, if anything, an even better candidate than the previous one for intellectual undecidability[47]. So, James' criterion applies, and Socrates cannot be labelled irrational for endorsing the Theory of Recollection, especially as there are good pragmatic grounds for accepting it: it will make us more vigorous inquirers, and so forth.

Finally, what should we say about the fact that James' conclusion is meant to apply to beliefs, whereas we are interested in hope? One possibility is to argue that hope is a weaker propositional attitude than belief anyway; so, if the belief that p cannot be criticised as irrational, and if p describes a state of affairs that is believed to be desirable, then, *a fortiori*,

[45] I discuss the applicability of James' conclusion to hopes shortly.

[46] Whether the option is intellectually undecidable in principle, or for all practical purposes is a difficult question, but I don't think we need to take a position on it here.

[47] Leaving out of the picture alleged proofs of the immortality of the soul.

it is not irrational to entertain the hope that p. However, I do not think that this is a good response, since what prompted this line of inquiry in the first place was a worry about the etiolated nature of some hopes. Having bought a lottery ticket, I may legitimately entertain the hope that I shall be a millionaire soon - as long, that is, as the expression of that hope involves little more than saying "I hope I win the lottery", and so on. Given that this was our concern at the end of section 2.3.1, it can hardly be correct to retreat to the position that hope is a very weak attitude to entertain, requiring little in the way of active commitment.

A better response is to argue that James' conclusion is best thought of as applying to hopes in the first place. Some of the examples that he discusses suggest that this might be the case. I, for one, have never been entirely convinced by the chasm jumping example, although clearly it describes a situation in which I must act in a certain way or perish. But, in making the attempt to jump the chasm, am I not just motivated by the *hope* that it will succeed? It could be argued that only full-blooded belief is enough to bring about the desired state of affairs; but, it seems to me that the situation James describes is so extreme, the stakes so high, that a suitably *fervent* hope would do the job just as well[48]. Consider also the possibility, even likelihood, that, if interviewed about my ordeal on a chat show, I will say "I really didn't believe I could make it". Obviously, I now *know* that such feats are not impossible, if the occasion demands them, but at the time, it is hard to see how a belief in such extraordinary abilities could form.

Further circumstantial evidence for reading James' conclusion in terms of hope rather than belief may be found in John Stuart Mill's short treatise *Theism*, in many ways an interesting precursor to James' essay[49]. Mill begins by subjecting the religious hypothesis to the critical scrutiny of scientific rationality, and concludes:

[48] I suggest shortly that the boundary between such a hope and belief is blurred anyway.

[49] Compare also this passage from Peirce's *Collected Papers* [2.113]: "When a hand at whist has reached the point at which each player has but three cards left, the one who has the lead often goes on the assumption that the cards are distributed in a certain way because it is only on this assumption that the odd trick can be saved. This is indisputably logical; and on a more critical analogous occasion there might be some *psychological* excuse, or even warrant, for a 'will to believe' that such was really the case. But all that logic warrants is a *hope* and not a belief. It must be admitted, however, that such hopes play a considerable part in logic. For example, when we discuss a vexed question, we *hope* that there is some ascertainable truth about it, and that the discussion is not to go on forever and to no purpose...". In this passage Peirce is clear that he prefers to use the word "hope", in circumstances where other philosophers might speak of a "will to believe". Note also that, as well as commenting upon the essential role played by hope in certain inquiries, this passage defends the legitimacy of their doing so.

> [T]he rational attitude of a thinking mind towards the supernatural, whether in natural or in revealed religion, is that of scepticism as distinguished from belief on the one hand, and from atheism on the other...[50]

In other words, the question of whether or not to believe in God has all the hallmarks of what James will later call an intellectually undecidable option. Therefore:

> The whole domain of the supernatural is thus removed from the region of Belief into that of simple *Hope*; and in that, for anything we can see, it is likely to remain[51].

However, Mill, like James, presses on:

> It is now to be considered whether the indulgence of hope, in a region of imagination merely, in which there is no prospect that any probable grounds of expectation will ever be obtained, is irrational, and ought to be discouraged as a departure from the rational principle of regulating our feelings as well as opinions strictly by the evidence[52].

Eventually, Mill proposes two general criteria for evaluating whether the indulgence of hope will undermine the proper "cultivation and regulation of the imagination": such indulgence must not disturb the "rectitude of the intellect"; and it must enhance life and elevate the character. In another anticipation of James' line of thought, Mill suggests that the relative weight accorded to these principles by different thinkers is likely to be decided by "individual temperament"[53].

So, there are two reasons to suppose that James' conclusion in "The Will to Believe" might really concern the legitimacy of acting on the basis of nothing more than a hope: at least one of the examples he describes seems best thought of as involving hope rather than belief; and Mill's argument, which closely parallels James', concludes that the idea of God is a legitimate object of hope, not belief. Even if these considerations do not establish that James should have called his paper "The Right to Hope", they do enough to make plausible the extension of James' conclusion to the kinds of hope we have been considering.

Finally, the converse point is also worth considering: it is not simply that the scope of James' conclusion may be extended to cover the kinds of hope we are interested in; it may be that

[50] J. S. Mill, *Theism*, Section V.
[51] *Ibid.*
[52] *Ibid.* Note that Mill's commitment to the rational regulation of the emotions (including hope) is quite general; so, as far as he is concerned, the case of religious faith should be no exception.
[53] *Ibid.*

the extraordinary nature of these hopes makes them, once they are taken as a basis for action, somewhat akin to beliefs[54]. Perhaps neither "hope" nor "belief" is appropriate to describe the propositional attitudes that occur in the unusual situations that interest us. My view is that it does not matter much how we choose to label these attitudes, as long as we are clear about what's going on. In all of the cases we have considered in this chapter - religious faith, scientific inquiry, James' friendship and chasm jumping examples - the issue is whether it is reasonable to make a commitment to a course of action, where the evidence that would normally justify this commitment is, for one reason or another, not available. James says enough, I believe, to establish that such a commitment is not unreasonable, given the momentousness of the issues involved. It is beginning to look as if the hope an inquirer places in the idea that the Universe is orderly, or in the Theory of Recollection, is not, after all, so misguided.

2.4 The Needs of Reason

The material in this section addresses the second concern identified at the end of section 2.3.1: that Aquinas' analysis of the relationship between hope and experience cannot readily be applied to the rather unusual hopes that we are interested in, which seem to have an especially tenuous connection to anything that might possibly be experienced. By way of contrast, Kant has a lot to say about the nature of such hopes, and it is to his writings that we now turn. In outline, the argument of this section is as follows:

1. Hope, especially hope in objects outside the field of possible experience, is a pervasive theme in Kant's writings, as brief reflection upon the nature of his critical project shows.

2. Analysis of some interesting passages in the early work, *Dreams of a Spirit Seer*, reveals that, on the whole, Kant views the relationship between hope and reason positively. However, not all hopes that have objects beyond the field of possible experience may reasonably be entertained.

3. The general question of the legitimacy of such hopes is answered in Kant's paper "What is Orientation in Thinking". Briefly, hope in an object that lies beyond the field of possible experience is reasonable, *as long as the hope answers to the needs of practical or theoretical reason.*

[54] This may be what Kant means when he insists that our conviction in the ideas that answer to the needs of practical reason is not inferior in kind to knowledge. See section 2.4, p72, for more details.

4. Kant's discussion of the needs of practical and theoretical reason, and the hopes that satisfy them, in the *Critique of Pure Reason* sheds further light on the nature and legitimacy of these unusual hopes.

We begin then with some general remarks about Kant's critical project, and the way in which hope emerges as an important strand within it. Towards the end of the *Critique of Pure Reason*, Kant claims that:

> All of the interests of my reason, speculative as well as practical, combine in the following three questions: *What can I know*?; *What ought I to do*?; and *What may I hope*[55].

Clearly, this statement assigns an important role to certain, yet to be specified, hopes in the life of reason. But from the pages that follow the above quotation, it is hard to determine what that role is, and the hopes that may legitimately fulfil it. In fact, it is difficult to find a single, extended treatment of this issue in any of Kant's writings; a puzzling state of affairs, given that both of the first two questions on the above list are the subjects of separate critiques [56]. This, I suggest, is because the theme of hope is so pervasive (albeit implicitly so) in Kant's writings that it rarely needs or receives explicit attention.

To see this, recall the way in which Kant sets the stage for his critical project, at the very beginning of the *Critique of Pure Reason*:

> Human reason has this peculiar fate that in one species of its knowledge it is burdened by questions which, as prescribed by the very nature of reason itself, it is not able to ignore, but which, as transcending all its powers, it is also not able to answer[57].

Kant goes on to say that, in its efforts to answer these questions, reason eventually finds itself "compelled to resort to principles which overstep all possible empirical employment", and thereby "precipitates itself into darkness and contradictions"[58]. If this is true, it is surely an understatement to describe the fate of human reason as nothing more than*peculiar*. If the nature of reason compels us to try to answer questions that transcend our intellectual powers,

[55]Kant, *Critique of Pure Reason*, A805. The passage is mentioned in the final chapter of A. W. Moore's, *The Infinite*.

[56] *The Critique of Pure Reason*, and *The Critique of Practical Reason*, respectively. Other works that address these questions include *Prolegomena to Any Future Metaphysics* and *Groundwork of the Metaphysic of Morals*. The third critique, *The Critique of Judgement*, is concerned mainly with aesthetic and teleological judgements, neither of which is obviously related to hope, and its relationship to reason.

[57] Kant, *Critique of Pure Reason*, A vii.

and if our attempts to answer these questions lead only to darkness and confusion, isn't the word *tragic* more appropriate? In such circumstances, one could be forgiven for succumbing to despair, and giving up on the philosophical enterprise[59]. Even Kant acknowledges that Metaphysics has become an object of ridicule, so that it resembles more a "matron outcast" than the "Queen of all sciences" it was once thought to be. And among the metaphysicians, a mood of indifference prevails, following the failure of dogmatists, sceptics and the "celebrated" Locke's "physiology of the human understanding" alike to solve the problems of philosophy.

Before giving in to despair however, we should recall our earlier discussion of the relationship between hope and despair[60], and, in particular, Gabriel Marcel's suggestion that "to trace the visage of despair is to conceive the features of hope". What this means is that wherever despair is a temptation, hope is also a viable option, as long as one adopts the right perspective on the situation – choosing, as it were, to see the glass as half full, rather than as half empty. This, I believe, is the strategy that Kant pursues. Immediately after his description of the prevailing mood in philosophy as one of indifferentism, we find this:

> [indifferentism is] the mother, in all sciences, of chaos and night, but happily [is] in this case the source, or at least the prelude, of their approaching reform and restoration. For it at least puts an end to that ill-applied industry which has rendered them thus dark, confused, and unserviceable[61]

Shortly afterwards, Kant again assigns a positive role to this mood of indifference by attributing it not to "levity", but to "the matured judgement of the age, which refuses to be any longer put off with illusory knowledge". So, although Kant acknowledges the defeatism that prevails in philosophical circles, he is not prepared to indulge in it himself. Like Socrates in the *Meno*, he views profound intellectual discomfort in a positive light: it is both a sign that error has been eliminated, and an opportunity to seek the truth with renewed vigour.

The way in which Kant pursues his inquiry is also reminiscent of our earlier discussion of the relationship between hope and despair. Taking his cue from the strides that had recently been made in mathematics and natural science he argues that:

[58] *Ibid.*

[59] Just as Meno does (or professes to do), when Socrates invites him to embark upon a joint inquiry into the nature of virtue.

[60] See above, pp44-46.

> Hitherto it has been assumed that all our knowledge must conform to objects. But all attempts to extend our knowledge of objects by establishing something in regard to them*a priori*... have... ended in failure. We must therefore make trial whether we may not have more success in the tasks of metaphysics, if we suppose that objects must conform to our knowledge[62].

This, of course, is Kant's "Copernican Revolution", so called because it turns one of the traditional questions of philosophy on its head, just as Copernicus found it necessary to replace the received view, that the spectator is at rest and the planets (apart from Earth) in motion, with its converse, in order to explain the apparent motion of the heavenly bodies. What is interesting about this strategy from our point of view is the resemblance it bears to the kind of change of aspect that is involved in the transition from despair to hope[63]. In effect, Kant suggests that we *can* extricate ourselves from a state of despondency regarding the state of metaphysics, by training ourselves to view the philosophical project in a new way; once we do this, we can engage in philosophical inquiry with a fresh sense of optimism. Socrates would have approved.

In this sense, Kant's critical enterprise may be seen as an attempt to rescue philosophy from the despair which, we might suppose, is its destiny; as such, hope, and the extent to which reason may avail itself of hope, is a pervasive theme in Kant's writings. In addition, there are a number of texts where Kant explicitly discusses the nature of hope - specifically, hope in things that lie beyond the field of possible experience - and its legitimacy. A useful starting point is the "pre-critical" work entitled *Dreams of a Spirit-Seer*[64]. The piece was published in 1766, fifteen years before the publication of the first critique. It was written in response to a letter from a friend, who wanted Kant's opinion on the claims of Swedenborg and his followers to have had direct experience of a realm of spirits. In the first half of *Dreams*, Kant conducts a methodical investigation into the possibility of such visions, considering firstly the case for them, and secondly the case against. In summarising his findings in the theoretical conclusion to the first half of the work, Kant introduces the analogy of the*scales of reason*. Just as ordinary scales are required by the civil law to be free from bias, and to accurately reflect the relative weights on either side, in the interests of fair trading, so does our own reason demand that it be free of bias when it weighs the relative merits of opposing viewpoints. To this end, Kant claims that he has eradicated every source of bias that may

[61] *Ibid.* A x.

[62] *Ibid.* B xvi.

[63] Especially if one buys into a Kuhnian account of the nature of scientific revolutions. See *The Structure of Scientific Revolutions*, chapter X.

[64] Published in Volume 1of *The Cambridge Edition of the Works of Immanuel Kant.*

have surreptitiously insinuated itself into his soul, and is now free of "blind attachments" and "prejudices".

Yet, however tenaciously Kant pursues the ideal of perfect balance in the scales of the understanding, the faculty of judgement, he acknowledges that there is one source of bias which it is particularly difficult to eliminate:

> But the scales of the understanding are not, after all, wholly impartial. One of the arms, which bears the inscription: *Hope for the future*, has a mechanical advantage; and that advantage has the effect that even weak reasons, when placed on the appropriate side of the scales, cause speculations, which are in themselves of greater weight, to rise on the other side[65].

In the above passage, Kant just reports the fact that considerations which chime with our hopes for the future are likely to weigh more heavily with us than ones which, objectively speaking, are of equal weight, but which do not enjoy this advantage. However, as the next two sentences suggest, Kant seems curiously ambivalent about how we ought to react to this state of affairs:

> This is the only defect, and it is one which I cannot easily eliminate. Indeed, it is a defect which I cannot even wish to eliminate[66].

In the first sentence, the bias of reason in favour of what accords with one's hopes is described unambiguously as a defect; made all the worse by its recalcitrance. However, the second sentence is more equivocal. It is most naturally read as claiming that it is *psychologically* impossible for me to eliminate this bias, but this is quite implausible. Granted, we may have a natural tendency to favour evidence that supports a view of the world as we would like it to be, and this tendency often serves the useful purpose of sustaining us in adverse circumstances. But surely, if we put our minds to it, most of us could learn to recognise when this tendency was influencing our judgement, and make allowances for it. Moreover, if we are sincere in our quest for the truth, then we just *do* wish to eliminate *all* intellectual bias, including that introduced by the influence of hope. A more plausible claim is that, as a matter of fact, it is *difficult* to wish to eliminate this bias, but this seems too banal to be what Kant has in mind, and, at any rate, it is not what he says. Is it possible, then, that Kant is making a much more interesting, normative claim about the relationship between hope and reason? Perhaps his point is that it would somehow be

[65] Kant, *Dreams of a Spirit Seer*; *Cambridge Edition of the Works of Immanuel Kant*, volume 1, p337.

unreasonable, to wish to eliminate the intellectual bias that hope introduces into the scales of reason.

If this is correct, Kant's attitude to hope is reminiscent of the views of Peirce and Plato regarding the role that certain hopes may reasonably play in our lives (*qua* rational inquirers), as already discussed in this and the previous chapter. To confirm this suspicion, we need to investigate why Kant might think that hope may, on occasion, legitimately influence the scales of reason. Some important clues may be found in one of his less well known essays, "What is Orientation in Thinking?"[67]. Examining these clues, and following up the issues they raise *via* some key passages in the *Critique of Pure Reason* will establish that Kant's views about the legitimacy of hope *are* remarkably similar to the views of Peirce and Plato, but are much more explicit about the peculiar nature of the hopes concerned.

"What is Orientation in Thinking?" appeared in 1786, between the publication of the first and second Critiques, and is an extremely useful introduction to Kant's critical system in its own right. Although the essay does not explicitly discuss the nature and legitimacy of hope, the related passages in the *Critique of Pure Reason*, mentioned in the previous paragraph, do. As we shall see, Kant thinks that the hopes he is interested in are rather unusual anyway, in that the degree of conviction we have in them makes them hard to distinguish from beliefs. In this respect, the hope theme is submerged, but nevertheless present, in "What is Orientation in Thinking?", in much the same way as it is in James' "The Will to Believe".

The question that Kant explicitly addresses in the essay, though, concerns what he calls the *speculative* use of reason; that is, the employment of reason outside the field of possible experience that is its rightful domain. The problem is that, if such a use of reason is to produce anything more than *idle* speculation, there must be some means of regulating it. But what could this be, given that, in its speculative employment, reason is not constrained by what it finds in the manifold of experience? Kant begins to answer this question by considering the claim of his friend, Moses Mendelssohn, that it is necessary to *orientate* oneself in the speculative use of reason by means of "common sense". Whilst Kant is suspicious of the idea that common sense can supply speculative reason with a means of orientation, believing that only reason itself can do *that*, he finds the orientation metaphor suggestive, and explores it at some length in the opening pages of the essay. Since we are

[66] *Ibid.*

[67] Collected in H. Reiss (ed.), *Kant: Political Writings.*

only interested in the clue this discussion provides regarding the proper use of speculative reason, I shall just give a brief summary of what he says.

In order to appreciate the implications of Mendelssohn's metaphor, Kant invites us to consider the case of *geographical* orientation[68], the kind of orientation we have when we can find our way around a given geographical region[69]. He points out that the proper meaning of "orientation" is "the use of a given direction in order to find the others"[70]. For example, I see the Sun rising, and, realising that that direction is East, am thereby able to locate the remaining points of the compass[71]. But, according to Kant, there is more to this process than meets the eye:

> For this purpose, however, I must necessarily be able to feel a difference within my own *subject*, namely that between my right and left hands[72]. I call this a *feeling* because these two sides display no perceptible difference as far as external intuition is concerned. If I were not able, in describing a circle, to distinguish between movement from left to right and movement from right to left without reference to any differences between objects within the circle... I would not know whether to locate West to the right or to the left of the Southernmost point of the horizon in order to complete the circle through North and East and so back to South[73]. Thus, in spite of all the objective data in the sky, I orientate myself *geographically* purely by means of a *subjective* distinction[74].

It does not matter, as far as we are concerned, that the last sentence in the above passage is incorrect, or, at best, ambiguous - Kant has already claimed that, in effect, orientation depends upon an *interplay* of objective and subjective factors, so it cannot depend solely on the latter, as the word "purely" suggests. What is important is that Kant believes he has found a clue to the nature of the orientation of speculative thought. It is how he develops this clue in the pages that follow that concerns us.

[68] He also discusses the case of, what he calls, *mathematical* orientation, but it is not clear how distinct this is from the geographical case. Rather than muddy the waters any further, I omit discussion of this case.

[69] See chapter 4 of this thesis for a detailed discussion of the nature of geographical orientation, and of how it relates to the second major theme of the thesis, understanding.

[70] *Ibid.* p238.

[71] The etymology of the term "orientation" reflects the usefulness of the direction of the rising Sun as a reference point: "orient" = "East".

[72] A Wittgensteinian might deny that our ability to distinguish left from right depended on any such feeling. But since Kant's discussion of the orientation metaphor is meant to be suggestive rather than conclusive, we need not worry too much about this. Our main concern, remember, is with how Kant goes on to apply the idea that there is a subjective element to orientation to the special case of the orientation of speculative thought.

[73] In Kant's example, the observer knows which direction is South to begin with.

[74] *Ibid.* pp238-9.

So, Kant claims to have established that the geographical orientation, which is relatively familiar, has both an objective and a subjective aspect. The next step, the difficult one, is to apply this insight to unfamiliar case of orientation in the realm of speculative thought - *logical* orientation, as Kant calls it. How can pure reason regulate itself when it tries to extend itself beyond the boundaries of possible experience, where there are no objects of intuition? Kant claims that orientation in this realm must again rely upon a subjective standard, which he suggests is a *feeling of need* that arises from the exercise of pure reason itself:

> It is possible to remain secure against all error if one does not venture to pass judgement in cases where one's knowledge is insufficient for the judgement in question.... But if it is not just a matter of indifference whether one wishes to make some definite judgement on something or not, if this judgement is made necessary by a real *need* (in fact by a need which reason imposes on itself), and if we are at the same time limited by lack of knowledge in respect of factors essential to the judgement, we require a maxim in the light of which this judgement can be passed; *for reason must sooner or later be satisfied*[75].

To illustrate the application of this principle, Kant returns to the theme of*Dreams of a Spirit Seer*. Reason does not *need* to posit the existence of a realm of invisible spirits in order to function properly. Hence, to speculate about the nature of spirits is described by Kant as a kind of inquisitiveness, which leads only to empty dreaming. Kant believes that the situation is quite different when we posit the idea of an original archetypal being, God, both as the supreme intelligence and the highest good[76]. In positing the existence of such a being genuine needs of reason - in both its practical and its theoretical employment - are satisfied.

Before looking at what these needs are, and how they are satisfied, let us examine the implicit contrast between needs and wants that informs Kant's discussion of logical orientation. What does this distinction amount to? One possibility is that the objects of our needs are, in some sense, *essential* to us, whereas things that we just want, however intense our desire, are not. There is a use of "needs" which, at first sight, seems to count against this idea, but which, when properly understood, helps us to appreciate it more clearly. I am thinking of a case like this: Genghis is burning the popcorn and I shout, "You need to turn the heat down!". This example seems to contradict our theory, since making popcorn can hardly be considered an essential activity. However, given that Genghis *is* trying to make

[75] *Ibid.* p240.
[76] *Ibid.* p241.

popcorn, he really does need to turn the heat down if he is to succeed. In this case, the need is *instrumental* to the achievement of a particular objective[77].

In addition to the needs something has in virtue of the goals it pursues, there are needs it has in virtue of the kind of thing it is. These are the things that are necessary in order for the *proper functioning* of the individual to continue[78], obvious examples being the need for food, warmth, and shelter. Both types of need involve the idea of a harmonious ordering of elements, be they actions or entities, either to realise a desired state of affairs, or to preserve an existing state. So, it looks as if one difference between needs and wants is that the former are constrained, if not determined, by the latter, but not *vice versa*.

Closely related to this point is the idea that wants reside solely in the subjective realm, whereas needs have an objective dimension. For example, if I say that I want a cream soda, generally speaking, other people have to take my word for it[79]. With needs, this situation is pretty much reversed. People are often pretty good at telling others what they need - if less adept at discerning their own real needs - and this practice is generally acceptable because, once an individual's wants are known, their needs are to a large extent publicly accessible, and subject to objective standards of evaluation. For example, given his desire for popcorn, Genghis' needs in the situation described earlier, depend upon such objectively specifiable facts as the temperature at which corn kernels begin to combust. And, in general, our needs depend upon our goals, and the properties of things which make them suitable for use as tools in the pursuit of those goals[80].

Let us return to the question of the orientation of thinking beyond the bounds of possible experience. Kant claims that such speculative use of reason is constrained by the requirement that it answer to a genuine need. If the preceding discussion of the distinction between needs and wants is correct, this means that recourse to speculative reason, if

[77] Another way of putting the point: the need in this example is tied to the giving of a reason; Genghis should turn the heat down *because* he doesn't want to burn the popcorn.

[78] This may be seen as a special case of the first type of need - if we assume that survival is an objective of most beings.

[79] This is not to deny that it is occasionally possible to tell someone what they "really" want, and to be right about this even if the individual concerned stubbornly refuses to admit it. However, these cases are exceptions, which show themselves only against a background of cases of the standard variety.

[80] So, how come Kant refers to needs as a *subjective* element, in his discussion of orientation in speculative thinking? I think the answer is that, from the perspective of the inquirer, the urge to go beyond the sphere of possible experience *is* experienced as something purely subjective, like a feeling. It is only when one adopts a third person philosophical perspective on this phenomenon that it's objective basis as a need of reason is revealed.

legitimate, must be an expression of the proper functioning of reason as it *normally* operates in the theoretical and practical spheres. Let us suppose that the goal of *theoretical* reason is to understand the physical universe, and that the goal of *practical* reason is to regulate conduct according to moral principles. According to Kant, we *need* to posit the existence of God, the supreme, archetypal being, if reason is to function properly in both of its principal aspects. Somewhat less remotely, and more relevant to our discussion of hope in the earlier part of this chapter, the idea that God exists is supposed to underwrite two intermediate assumptions:

1. As far as theoretical reason is concerned, the idea of God as the supreme intelligence, designer and architect of all creation, underwrites the idea that the universe is essentially orderly.

2. Practical reason, on the other hand, views God in his other aspect, as the judge of the moral worth of his creatures, and posits the idea of an intelligible moral order.

Although we are mainly interested in theoretical reason, it will prove illuminating to examine how Kant applies these ideas to the case of practical reason first. As stated above, the goal of practical reason is the regulation of conduct; and it needs to posit the idea of an intelligible moral order if it is to function properly. The moral order that Kant envisages is not important, as far as we are concerned, but it is worth describing in a little detail to see how the needs of reason impose themselves in the practical sphere.

In its most well known form, Kant's Categorical Imperative instructs us to "act only according to that maxim whereby you can at the same time will that it should become a universal law"[81]. This declares an action to be necessary without reference to any purpose, in contrast to the *hypothetical* imperatives that govern most of our behaviour[82]. Kant intends the Categorical Imperative to have universal application, regardless of our purposes and desires, since it is a principle of duty whose necessity is guaranteed by reason itself. The problem is that nothing Kant says rules out the possibility of a state of affairs where the virtuous (those who do their duty) are miserable, and the wicked are happy. Indeed, cynics might say that this is exactly what does happen, "in the real world". Kant, sees this as unsatisfactory from the point of view of practical reason itself: although the virtuous man acts out of duty with

[81] Kant, *Groundwork of the Metaphysics of Morals*, p421.

[82] To recycle a previous example, *turn the heat down if you don't want to burn the popcorn.*

no expectation of a reward[83], it still runs counter to reason that he should be unhappy as a result of his obedience to the Moral Law. The lack of comensurability between virtue and happiness in this world appears to mock the pretensions of practical reason. Hence this very same reason demands that we postulate an intelligible moral order in which happiness is apportioned according to virtue by an all-powerful, benevolent ruler. As this is obviously not the present state of affairs, the realisation of this order must lie in the (possibly distant) future; hence it is also necessary to entertain the idea of a soul which is immortal. Here is Kant's summary:

> Morality by itself constitutes a system. Happiness, however, does not do so, save in so far as it is distributed in exact proportion to morality. But this is possible in the intelligible world under a wise Author and Ruler. Such a Ruler, together with a life in such a world, which we must regard as a future world, reason finds itself constrained to assume; otherwise it would have to regard the moral laws as empty figments of the brain...[84]

What attitude are we permitted to entertain towards this idea of a perfect moral order in a future life? In the above passage Kant implies that it is an assumption, albeit a necessary one. This is a rather unspecific term, but later on, Kant clarifies his meaning:

> Morality, taken by itself, and with it, the mere *worthiness* to be happy, is also far from being the complete good. To make the good complete, he who behaves in such a manner as not to be unworthy of happiness must be able to *hope* that he will participate in happiness[85].

This idea - that the intelligible moral order presupposed by the nature of practical reason is a legitimate object of *hope* - ties in with our definition from section 2.2:

S hopes that p iff:

1. S desires that p *or* S believes that p is good.
2. S believes that p is uncertain.
3. S believes that p is possible *or* has no reason to believe that p is impossible.

Firstly, from the standpoint of practical reason it is highly desirable that happiness is proportionate to virtue, though this desire may not influence the moral will. Secondly, the

[83] Of course, entertaining such desires is permissible, even inevitable, given our limitations; but allowing them to influence the will - doing one's duty *because* one desires happiness - is another matter.

[84] Kant, *Critique of Pure Reason*, A811.

[85] *Ibid.* A813.

realisation of the *summum bonum*, as Kant calls the intelligible moral order he describes, is far from certain. And thirdly, the idea of the *summum bonum* does not involve a contradiction, so it must be a logical possibility; and there is no reason to think that it is an empirical impossibility. In sum, the idea of an intelligible moral order is something that we may legitimately hope for since it answers to a genuine need of reason.

The above discussion of the needs of practical reason is taken from the *Critique of Pure Reason*. The discussion of the same issue in "What is Orientation in Thinking?" goes further, and returns us to the needs of theoretical reason in the process:

> Much more important [than the need of reason in its theoretical use], however, is the need of reason in its practical use, because this is unconditional, and because we are compelled to assume that God exists not only if we *wish* to pass judgement but because we *must pass judgement*[86].
>
> On the other hand, a *rational belief* that is based on the need to use reason for *practical* purposes could be described as a *postulate* of reason; for although it is not an insight capable of fulfilling all the logical requirements for certainty, this conviction of truth is not inferior in degree to knowledge... even if it is totally different from it in kind[87].

In these passages, Kant *contrasts* the needs of practical and theoretical reason, attributing to the former a kind of certainty, not granted to the latter. The remarks are difficult to interpret, but the underlying idea seems to be that refraining from action, and the practical deliberation that goes with it, just isn't an option[88], whereas, *prima facie*, it is possible to refrain from theoretical speculation. Hence, the suggestion that the employment of theoretical reason is conditional - it only comes into play *if* we wish to understand the physical universe - in a way that the use of practical reason is not. If correct, this bestows upon the idea of an intelligible moral order a kind of *unshakeability*; it is an idea I will not relinquish under any circumstances, despite the fact that it is, strictly speaking, only a hope.

This is an interesting idea, all the more so, from our point of view, as I believe that Kant exaggerates the difference between the needs of practical and theoretical reason. An alternative point of view, which highlights the pervasiveness of theoretical reason in our lives, may be found in the following passage from Aristotle:

[86] Reiss (ed.) *Kant: Political Writings*, p242.
[87] *Ibid.* p245.

> All men by nature desire to know. An indication of this is that delight we take in our senses; for even apart from their usefulness they are loved for themselves; and above all others the sense of sight. For not only with a view to action, but even when we are not going to do anything, we prefer sight to almost everything else. The reason is that this, most of all the senses, makes us know and brings to light many differences between things[89].

Aristotle ascribes the origin of this desire to know to a sense of wonder:

> For human beings originally began philosophy, as they do now, because of wonder, at first because they wondered at the strange things in front of them, and later because, advancing little by little, they found greater things puzzling[90].

Again, the search for understanding that results is not driven by the need to solve practical problems:

> For it was only when practically everything required for necessities and for ease and leisure pursuits was supplied that they began to seek this sort of understanding; clearly, then, we do not seek it for some further use[91].

The conception of the place of theoretical reason in human life set out in the above passages is, to me, at least as plausible as Kant's. It *may* be true, a Kant suggests, that the theoretical employment of reason is not inevitable. Still, a life devoid of all curiosity about the workings of the natural world does not sound like a recognisably human life. As Aristotle suggests, we just seem to be naturally disposed to find certain things wonderful, and thence, worthy of explanation. This does not mean that everyone has an irresistible urge to attempt to solve the riddles of fundamental physics; but it does mean that wonder and curiosity have a way of insinuating themselves into all but the most impoverished of lives. So, although Kant is probably correct to insist that the employment of theoretical reason is conditional, as the employment of practical reason is not, Aristotle reminds us that, as a matter of fact, it is human nature to inquire into the workings of the physical universe: in other words, the way of rational inquiry is not something that we can just leave on a whim.

This has a bearing on the worry, raised at the end of chapter 1, that resting the possibility of rational inquiry on mere hopes, as Peirce does, is not enough: why not just give up on the

[88] Even opting to spend the whole of one's life in a sensory deprivation tank in the middle of a desert is a practical decision

[89] Aristotle, *Metaphysics* I.i, 980a 21-27. See also J. Lear, *Aristotle: The Desire to Understand*, chapter 1.

[90] *Ibid.* I.i 982a 11-15.

[91] *Ibid.* I.i 982a 22-25.

whole scientific enterprise and do something else? Aristotle's suggestion that the search for understanding is not something we can just turn our backs on (without sacrificing an important aspect of our humanity) goes some way towards answering this question. We should also point out that it is somewhat unjust to refer to the hopes that underpin the possibility of rational inquiry as "mere hopes", given the central role that theoretical reason plays in most of our lives. Though these hopes may not enjoy the "conviction of truth... not inferior in degree to knowledge" that the hope in an intelligible moral order does, they cannot fall far short. This, I believe, is as good a response to our worry as we can give. If it still seems to leave us in a potentially tragic predicament - compelled by nature to search for order, with no absolute guarantee that it exists, or that, if it does, our minds are capable of grasping it - perhaps we should recall Socrates' words of encouragement: there is no need to despair; an attitude of fortitude and optimism is just as viable, and probably more rational, from a practical standpoint. This is what we should seek to cultivate as we make our way along the path of inquiry.

Although that is all I really want to say about Kant, I shall, for the sake of completeness, describe what he believes needs to be posited as an object of hope, from the standpoint of theoretical reason. The essential idea is of an ideal physical order, the equivalent of the intelligible moral order that practical reason is compelled to posit. Kant's most detailed account of what this idea entails may be found in the "Regulative Employment of the Pure Ideas of Reason" of the *Critique of Pure Reason*[92]. Briefly, there are supposed to be three methodological principles for the regulation of scientific practice[93]: the principle of *homogeneity*; the principle of *variety*; and the principle of *affinity*.

Kant sometimes refers to the principle of homogeneity as the principle of*genera*. The basic idea is that the scientific inquirer should aim for*unity* in the concepts she uses to understand the world. Stated as a formal methodological principle:

> Develop a conceptual structure that will reduce the complexity of empirical knowledge by searching for generic concepts and laws of which known empirical concepts and laws will be specifications[94].

[92] Kant, *Critique of Pure Reason*, A642-68.

[93] In fact, it is clear at the outset that these must be more than just methodological; I return to the issue of the status of the principles after spelling out their content.

[94] Here, and in the following passages, I use T. A. Wartenberg's formulations. See his "Reason and the Practice of Science" in the *Cambridge Companion to Kant*.

Maxwell was theorising in accordance with this maxim when he derived the electromagnetic field equations that brought together, the realms of electricity and magnetism. Likewise, contemporary theoretical physicists entertain hopes of a grand unified theory of the four fundamental forces of nature.

The principle of variety complements the search for unity in experience. As its name suggests, the principle instructs us to look for variety where previously we might have been aware only of similarity. Alluding to Ockham's razor, Kant states this principle in Latin: *entium varietates non temere esse minuendas*; the variety of entities is not to be thoughtlessly reduced[95]. A good example of this principle in action is the discovery earlier this century that certain elements exist in two or more forms, known as isotopes, depending upon the number of neutrons in the nucleus. Regulating our inquiries according to the principle of variety ensures that our empirical knowledge of the world is as comprehensive as it can be, and, in the limit, that the final system of scientific knowledge is *complete*[96].

The third principle, that of affinity, is rather more difficult to describe. Roughly speaking, it instructs us to look for *continuity* in our experience of the empirical world, to, as it were, fill in the holes in our conceptual net. As an example, Kant gives the formulation of the hypothesis that planetary orbits are ellipses when it was discovered that the observations did not support the longstanding belief that they were perfect circles:

> Thus, for instance, if at first our imperfect experience leads us to regard the orbits of the planets as circular, and if we subsequently detect deviations therefrom, we trace the deviations to that which can change the circle, in accordance with a fixed law, through all the infinite intermediate degrees, into one of these divergent orbits... and thus we come upon the idea of an ellipse[97].

According to Kant, the principle of affinity is a kind of synthesis of the principles of homogeneity and variety, for "only through the processes of ascending to the higher genera and of descending to the lower species do we obtain the idea of systematic connection in its completeness"[98]. Just as nature is said to abhor a vacuum theoretical reason abhors the thought of fundamental discontinuities between species within the same genera.

[95] *Ibid.* A656.

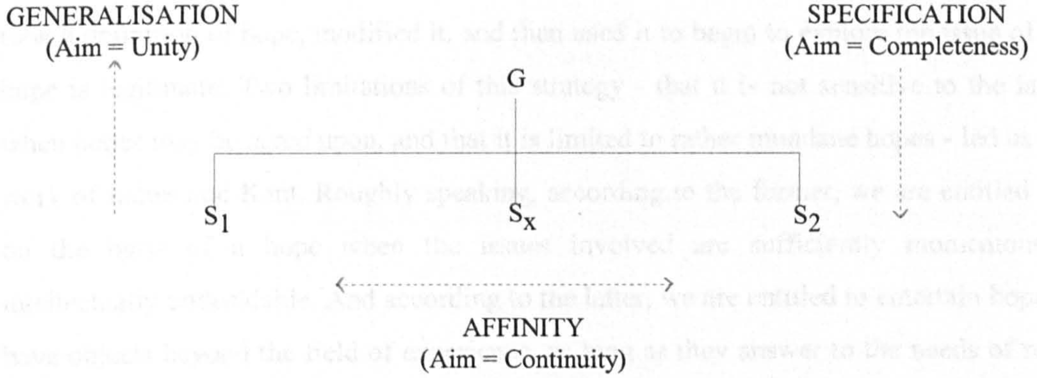
[96] Setting aside questions about whether such a "final" system is humanly attainable.

[97] Kant, *Critique of Pure Reason*, A662.

[98] *Ibid.* A658.

In sum, three methodological principles regulate the theoretical use of reason, the relationships between them being as illustrated in the diagram below:

Figure 2.3: Methodological Principles of Theoretical Reason



As might be predicted by now, Kant is not content to let generality, specificity and affinity remain methodological principles. For it would not make sense for scientists to conduct their enquiries in accordance with the principles if there were not also, what Kant calls, a *transcendental* way of understanding them, one which alludes to objects beyond the realm of possible experience. With reference to the principle of homogeneity, for example, Kant says:

> For with what right can reason, in its logical employment, call upon us to treat the multiplicity of powers exhibited in nature as simply a disguised unity... if it be free to admit as likewise possible that all powers may be heterogeneous, and that such systematic unity of derivation may not be in conformity with nature?... In order therefore to secure an empirical criterion *we have no option save to presuppose the systematic unity of nature as objectively valid and necessary*[99].

Hence, the idea of a unified, complete, and continuous physical order is a presupposition of theoretical reason, something which it is necessary to assume in order for inquiry to make sense. However, Kant makes clear we cannot *know* that the natural world accords with this idea[100]; but again, it is something that we may legitimately hope for, as answering to a genuine need of reason.

[99] *Ibid.* A651

[100] So Wartenberg goes too far in claiming that the principles express privileged items of transcendental *knowledge*.

2.5 Summary

This chapter has been something of a patchwork. My aim has been to assemble a coherent account of the nature and legitimacy of hope in general, and the hopes that underpin the possibility of rational inquiry in particular, by stitching together ideas taken from the works of several philosophers, principally Aquinas, William James, and Kant. From Aquinas, I took a definition of hope, modified it, and then used it to begin to explore the issue of when hope is legitimate. Two limitations of this strategy - that it is not sensitive to the issue of when hopes may be *acted* upon, and that it is limited to rather mundane hopes - led us to the work of James and Kant. Roughly speaking, according to the former, we are entitled to act on the basis of a hope when the issues involved are sufficiently momentous, yet intellectually undecidable. And according to the latter, we are entitled to entertain hopes that have objects beyond the field of experience, as long as they answer to the needs of reason. The hopes that sustain the possibility of rational inquiry meet both of these criteria, and may therefore be legitimately entertained, and acted upon.

Chapter 3: Understanding, Models, and the *Meno*

3.1 Introduction

In chapter 1, I suggested that two of the key themes of the *Meno* dialogue are:

- Understanding is the goal of inquiry, or at least of many of the inquiries that we hold in the highest regard.

- The possibility of rational inquiry depends upon our entertaining hopes in things that are not amenable to rational demonstration.

Chapter 2 examined the second of these themes in detail, and arrived at some general conclusions about the nature and legitimacy of hope in things that are not amenable to rational demonstration. The remaining chapters of the thesis, beginning with this one, will explore the first of the above themes in comparable detail, building on the following ideas about the nature of understanding discussed briefly in chapter 1:

- Understanding is, at least partly, a result of trying to make sense of new information by seeing how it squares with prior beliefs and intuitions.

- When one has understanding of a given domain, one's beliefs enjoy a distinctive and valuable form of stability.

Our inquiry into the nature of understanding will concentrate, for the most part, upon the second of these claims, although the first receives some attention in section 3.4 of this chapter. Before we can sensibly discuss either of them, however, we need to make a couple of important distinctions. This is the main purpose of section 3.2, which, as well as discussing the familiar distinction between linguistic and non-linguistic understanding, introduces a less familiar distinction between what I call *dynamic* and *static* understanding. The dynamic and static aspects of understanding are related in that we may reasonably hope that the former will, in the long term, lead to the latter. Section 3.3 completes the preliminaries by returning to the passages in the *Meno* that led us to suppose that the goal of inquiry is understanding, in search of further clues about the nature of understanding. I begin to develop these hints in section 3.4, which gives an account of the difference between what I call *minimal*, and *non-minimal comprehension*; the ideas in this section tell us something about the nature of *dynamic* understanding. In section 3.5, I begin to develop an account of *static* understanding, suggesting that it is a matter of having an adequate*model* of its object; this idea is the topic of chapters 4 and 5.

3.2 Understanding: Preliminary Issues

It is often claimed that there is an intuitive contrast between knowing lots of distinct facts about a given domain and having genuine understanding of it. The latter, it is said, is a holistic phenomenon, preferable to the former, on the grounds that it somehow demands a deeper level of familiarity with its subject: anybody, after all, can memorise a load of facts, but it takes effort and insight to truly understand a given domain. If true, this might account for why understanding appears to be held in such high regard - as the goal of inquiry, no less - in the *Meno* dialogue[1].

Although the intuitions described in the previous paragraph are crude[2], I believe that they*do* point in the right direction. Furthermore, I believe that the distinction between knowledge and understanding that these intuitions attempt to articulate is an important one[3]. But it is also vague, its precise nature having proven notoriously difficult to pin down. In the next three chapters, I aim to remedy this situation.

It is useful, if a little daunting, to begin by reminding ourselves of the number and variety of things that may be properly described as potential objects of understanding. A partial list would include the following:

- Theories
- Utterances
- Situations

[1] If I am correct in thinking that *episteme*, the goal of inquiry in the *Meno*, is best translated as "understanding". See section 3.3 for further details.

[2] Is it really correct, for instance, to think in terms of a *contrast* between understanding, and knowing lots of facts? Isn't knowing a lot of facts about a given domain usually at least *prima facie* evidence of understanding? For this reason, and possibly others besides, some readers might prefer to regard understanding as a special kind of knowledge, or to think in terms of a distinction between knowledge with, and knowledge without understanding. For our purposes, it does not matter how we choose to label the distinction, as long as its existence, at an intuitive level, is acknowledged.

[3] Though surprisingly few philosophers have paid much attention to it. Exceptions include: N. Cooper, in a series of papers; A. W. Moore, in "Ineffability and Reflection: An Outline of the Concept of Knowledge"; J. M. Moravcsik, in "Understanding"; R. L. Franklin, in "Knowledge, Belief and Understanding"; and J. Rosenberg, in "On Understanding the Difficulty in Understanding Understanding". See also L. Zagzebski, *Virtues of the Mind*, pp43-51.

- Poems
- People
- Languages
- Pieces of music
- Events
- Ranges of phenomena

The list is bewildering at first sight, but I think that we can impose some order on it by drawing two distinctions. Firstly, items on the list may be classified as either essentially linguistic or essentially non-linguistic. The former category includes utterances, languages, poems and (perhaps) theories[4]; and the latter includes people, events, situations and ranges of phenomena. Most philosophical discussions of understanding concentrate on its linguistic manifestation, and, perhaps, tend to exaggerate the distance between it and its non-linguistic counterpart. For what it's worth, my intuition is that this may be a mistake, the distinction being more a matter of degree than anything else: linguistic phenomena are just some of the things in the world that it is possible to seek, and attain, understanding of. At any rate, this thesis is mainly concerned with those cases in which the object of understanding is essentially non-linguistic[5], and makes no assumptions about the relationship between linguistic and non-linguistic understanding[6].

The second distinction is more important as far as we are concerned: it is the distinction between what might be called the *dynamic* and the *static* aspects of understanding. This distinction emerges most clearly in the linguistic case, where people may be said to understand not only individual utterances in a given language, but also the language itself. In the first sense, understanding is something that occurs at a specific time, sometimes as a

[4] Or perhaps not! The discussion of the syntactic and semantic conceptions of scientific theories in chapter 5 has a bearing upon this point.

[5] Though this does not preclude consideration - in section 3.4, for example - of cases where one learns about the object of understanding by means of language.

[6] A further complication: I may understand all the sentences in a given text, without understanding what their author was trying to achieve (why she thought those sentences were interesting, how the premises of her arguments were meant to support her conclusions, and so forth). Arguably, what I lack in this case is non-linguistic understanding of a linguistic object. This suggests that the distinction between linguistic and non-linguistic understanding may not be the same as the distinction between understanding linguistic objects and understanding non-linguistic objects, after all. As we are concerned mainly with non-linguistic understanding of non-linguistic objects in this thesis, we can safely set aside this point.

result of active effort, but usually, at least in the linguistic case, not. In the second sense, understanding is a *state* of an individual that persists over time[7,8].

Both kinds of understanding also occur in the non-linguistic case. At this stage, I do not wish to delve into the nature of static understanding of a non-linguistic domain, as that is the topic of chapters 4 and 5. In the meantime, it is safe to assume that a professor of atomic physics, say, has as good a grasp of that particular discipline as anyone does. This grasp of her chosen subject, which, *very* roughly speaking, amounts to the ability to solve many problems about atomic physics, is what I call static understanding in the previous paragraph, as it is something that persists in the professor over a long period of time. Dynamic understanding, in this context, might occur during a tutorial with a student: having tried to explain some subtle feature of the behaviour of hydrogen atoms in a number of different ways without success, the professor finally hits upon the right approach, one that the student can relate to; in these circumstances, it is appropriate for the student to say something like, "*Now* I understand", signifying that an episode of dynamic understanding has just taken place[9].

It would be surprising if these two aspects of understanding were entirely distinct, and, in fact, they are not. For, as a rule, it is via episodes of dynamic understanding that we come to build up understanding in the static sense. For example, when learning a language, it is through grasping the significance of many utterances that we come to acquire mastery of the language itself, the ability to generate indefinitely many utterances in future. Likewise, to return to our non-linguistic example, we grasp the significance of a particular fact about atomic physics when we integrate it with the rest of our knowledge about that, and related domains. And, if we pursue our inquiries diligently, we may reasonably expect that our understanding of atomic physics will grow, as manifested in our ability to answer more and more questions about it.

The idea that static understanding is arrived at via a series of episodes of dynamic understanding sounds quite plausible, but there are a couple of complications to bear in

[7] Though, of course, it need not, and typically does not, persist indefinitely. In particular, understanding, in this sense, may diminish over time if it is not exercised regularly. My present understanding of French, for example, is much weaker than it was just prior to taking my O levels.

[8] An alternative way of describing this distinction, in terms that will be explained later, is that, dynamic understanding is exercised in the construction and modification of a model, whereas static understanding is a matter of having and using a model.

[9] Of course, merely saying this is not sufficient to *demonstrate* understanding; to do that, the student's claim to understand must be tested by further questioning. See sections 3.3 and 3.4 of this chapter for further discussion of this point.

mind. Firstly, it seems possible that there could be an atomic physics pill, say, which, if swallowed, gives an individual understanding of the subject without the arduous study that this normally requires; this shows that episodes of dynamic understanding are not necessary for static understanding to arise.

In many ways, I am content to reply that this is not the way that static understanding occurs in reality. However, it is worth briefly considering a second case which, whilst not as outlandish as the atomic physics pill, remains (for now, at least) within the realms of science fiction, as this will cast doubt on whether the first case represents a genuine possibility. Suppose that understanding of atomic physics could be acquired, not by taking a pill, but by plugging some kind of chip into a socket in one's head; again, this appears to be an example of static understanding arising without intervening episodes of dynamic understanding. But there is a subtle difference between the atomic physics pill and the atomic physics chip: it is difficult to imagine that the latter could exist without*someone*, at some point, having taken the trouble to acquire understanding of atomic physics through their own hard work[10]. If this intuition is correct, the static understanding that is acquired when one plugs the chip in is, ultimately, parasitic upon series of episodes of dynamic understanding after all. Although this is not conclusive proof that the atomic physics pill is not really possible, it should at least make us wary of accepting the example at face value.

The second complication is that sometimes an episode of dynamic understanding may force us to acknowledge that we do not understand a given domain as well as we thought we did. As I suggest in the next section, this is what happens to Meno in the course of the dialogue, as first one, then another, of his attempts to define virtue succumb to Socrates' critique. How does this square with the idea that episodes of dynamic understanding lead to static understanding? The best way to deal with this worry, I believe is to draw on some of the ideas from the first two chapters of the thesis. In particular, it may be that, although an episode of dynamic understanding occasionally makes us feel that we understand (in the static sense) less than we thought we did, this is usually a temporary situation. For, as in the *Meno*, the very perplexity that this situation creates, is itself a stimulus to further, more rigorous inquiry, with the associated prospect of further, more productive, episodes of dynamic understanding later on. And, to return to another theme of the thesis, even though we cannot be certain of it, we may surely *hope* that the outcome of all of this intellectual

[10] This relates to the discussion of the difference between models and look up tables in section 3.5; see pp111-115 for further details.

effort will be an improved understanding of the domain concerned. In short, the two aspects of understanding we have identified fit neatly into the conception of inquiry implicit in the *Meno* and defended explicitly by Peirce[11].

3.3 Understanding in the *Meno*

Let us begin our inquiry into the nature of understanding by returning to the passages in the *Meno* discussed in chapter 1, in search of clues. Recall that in English translations of the *Meno*, the term *episteme*, which describes the goal of rational inquiry, is usually rendered as "knowledge". However, in chapter 1, we saw that several passages in the dialogue suggest that this is a mistake[12]. Consider once more the conclusion of Socrates' dialogue with Meno's slave boy, Nemo[13]. The slave boy has correctly answered Socrates' questions, and has thereby been led to the solution to the geometrical problem posed at the start of the conversation. At this point, a question arises: "Can Nemo be said to *know* that the square constructed upon the diagonal has twice the area of the original square?". Socrates appears to think not:

> At present. these opinions, being newly aroused, have a dream-like quality. But if the same questions are put to him on many occasions and in different ways, you can see that in the end he will have a knowledge on the subject as accurate as anybody's[14].

This remark may be read in at least two different ways, though this turns out not to matter, as far as we are concerned:

1. Nemo does have knowledge, but it is not as "accurate" as that of someone who has been asked the same questions "on many occasions and in different ways".
2. Nemo does not yet have knowledge, but will acquire it if the same questions are put to him "on many occasions and in different ways".

[11] See chapter 1 of this thesis for more on the nature of rational inquiry, according to Plato and Peirce.

[12] I am not the first to pick up on this point. M. Burnyeat notices it in "Socrates and the Jury: Paradoxes in Plato's Distinction Between Knowledge and True Belief". "Wittgenstein and Augustine *de Magistro*", and "Aristotle on Understanding Knowledge" by the same author explore similar territory.

[13] The dialogue with the slave runs from *Meno* 82b to 85b.

[14] Plato, *Meno*, 85c.

Since Socrates does not exactly *say* that Nemo lacks knowledge, the first reading seems the more faithful. However, as we saw in chapter 1, Socrates later contrasts *episteme* with true belief in the following terms:

> True opinions are a fine thing, and do all sorts of good so long as they stay in their place; but they will not stay long. They run away from a man's mind, so they are not worth much until you tether them by working out the reason. That process, my dear Meno, is recollection, as we agreed earlier. Once they are tied down they become *episteme*, and are stable[15].

Although both true belief and *episteme* are equally useful for practical purposes, in that a course of action informed by a the former will fare no worse than one informed by the latter, *episteme* is more valuable because, unlike true belief, it is "tethered" by reason, and hence is not so apt to disappear like the legendary statues of Daedalus[16]. These remarks support the second interpretation of the above passage, for in it, Socrates is concerned about the "dream-like" quality of Nemo's newly aroused opinions. If, as I suggested in chapter 1, the salient feature of dreams is their evanescence, their tendency to slip our minds upon waking, then Socrates could be claiming that the slave boy has a true opinion but does not yet have *episteme* in the above passage.

However, as I remarked earlier, it does not matter which of the above readings is correct because, on either of them, the slave boy's education in elementary geometry is not yet complete. On top of this, either reading leaves us with an interpretative problem. The first implies that knowledge can be more or less accurate. But if S knows that p, then, according to just about any analysis of knowledge, p is true, and S' knowledge is therefore as accurate as it ever can be. The second reading denies that Nemo has knowledge. But again, if anything like the tripartite analysis of knowledge is correct, he does: he has a true belief which is justified (by the geometrical construction that Socrates draws in the sand). Either way, the idea that *episteme* is best translated as "knowledge" looks suspect.

In contrast, the idea that *episteme* may best be translated as "understanding" begins to look appealing. Firstly, understanding is something that admits of degrees in a way that knowledge does not. True, we do not normally talk about understanding being more or less *accurate*; but we do think that understanding can be more or less complete, or more or less

[15] *Ibid*. 98a
[16] *Ibid*. 97d.

correct, which is probably enough for our purposes[17]. Secondly, having a justified true belief, or even a set of justified true beliefs is not sufficient for understanding the object of those beliefs. One way of illustrating this point, relevant to what comes later[18], is that a student who just repeats, parrot-fashion, whatever the professor of atomic physics says about a given question, does not thereby exhibit understanding[19]. Yet, assuming the professor expresses her beliefs sincerely, the student can acquire any number of true beliefs[20] by this method. Furthermore, the student will be justified in holding beliefs acquired in this fashion since they emanate from an authoritative source.

So, encouraged by the above results, let us pursue the idea that *episteme* is best translated as "understanding" by seeing what else Socrates says about the slave boy's cognitive state at the end of his dialogue with him. In particular, let us consider what more, according to Socrates, Nemo must do in order to convert his true opinions into *episteme*[21]. In the passage with which we began this section, Socrates appears to make two distinct suggestions:

A. Nemo must answer the same questions correctly when they are put to him *on many separate occasions*.

B. Nemo must answer the same questions correctly when they are put to him *in different ways*.

Some commentators on the *Meno* emphasise the first of these at the expense of the second[22]: if what makes true opinion inferior to *episteme* is that it is apt to slip one's mind, what better way of "tethering" it than to drill it into the student? However, there are several reasons to believe that this emphasis is misplaced. Firstly, the observation that repetition is sometimes a causal condition of someone coming to acquire, and subsequently recall, a justified true belief is of little philosophical interest. Secondly, quite apart from its philosophical inadequacy, the idea that education boils down to nothing more than rote learning is extremely crude from a pedagogical point of view. Given that these points seem so obvious,

[17] See sections 4.3 and 4.4 for more ways in which understanding might be said to come in degrees.

[18] See section 3.5, p116.

[19] I take it that this is fairly intuitive.

[20] I assume that the student is at least familiar enough with the subject to understand what most of the terms used mean - so she is not simply, as it were, copying the noises the professor makes.

[21] For convenience, I shall assume from now on that the second reading is correct: Nemo has true opinions, but does not yet have *episteme*.

[22] I think that Lawrence Powers falls into this trap in his otherwise illuminating discussion of the *Meno*, "Knowledge by Deduction".

we ought to invoke the principle of charity, and look for a richer interpretation of what Socrates says. So, let us examine the second possibility in more detail, and, in particular, consider what posing the same questions "in different ways" might mean.

One possibility is that the same questions are asked, but their linguistic expression is altered. On this interpretation of B, for example, all of the following would count as different ways of asking the same question:

- What is the sum of 2 and 3?
- 2 plus 3 makes what?
- What do you get if you add 2 to 3?

Note how slight the changes involved are though. Indeed, the changes are *so* superficial that I don't think this possibility can be what Socrates has in mind. Remember our concern about A above was that it places undue emphasis on repetition and rote learning. This prompted us to look for another way of understanding the process whereby true belief is converted into *episteme*. But it is hard to see how *this* anxiety is dispelled merely by insisting that minor grammatical changes be introduced each time the question is asked: the transformation is surely too slight to make any difference. In effect, this unsophisticated reading of alternative B reduces it to the previously rejected alternative A. I conclude that we must look for another way of interpreting the phrase, "the same questions in different ways".

So, suppose that, during an elementary mathematics lesson, Nemo and his master, Meno, are informed that:

- 2 + 3 = 5.
- The two lowest prime numbers are 2 and 3.

A teacher interested in how well they have followed the lesson might put the following questions to them:

- What is the sum of 2 and 3?
- What are the two lowest prime numbers?

Assuming that neither pupil has forgotten, both will respond with two correct answers. Note that one way of increasing the likelihood of this outcome would be to ask Nemo and Meno the same questions over and over again, correcting them each time they make a mistake. In other words, if all the teacher is interested in is getting Nemo and Meno to accurately repeat what she has told them, drill is sufficient.

Now suppose instead that the teacher asks, "What is the sum of the two lowest primes?". Even though she has not *told* Nemo and Meno the answer to this question, it is reasonable to expect them to be able to provide it, given the information they have at their disposal. Furthermore, it seems fair to say that this is the same question as, "What is the sum of 2 and 3?"; we have just, as Socrates might say, put it to them in a different way, one that both students are equally qualified to recognise, given the previous content of their lesson. Finally, if only Nemo, say, can answer this question correctly, there is a strong inclination to say that he is the better student. For he has succeeded in putting two pieces of information learnt during the lesson together, whilst the same two pieces of information remain stubbornly separate in Meno's mind. The most natural way of describing this situation is to say that Nemo has shown greater *understanding* of the material than his master, Meno.

This example, I believe, illustrates what Socrates thinks is necessary to convert true opinion into *episteme*. Roughly speaking, the requirement is that questions are put to the student that have not been asked explicitly before, but that are nevertheless answerable, given other information at the student's disposal. To return to our example, the question is designed to get the student to recognise that "the two lowest prime numbers" and "2 and 3" are co-referential terms, and to substitute the latter for the former in the question posed to obtain a question she can answer. We may speculate that the effect of this procedure, if repeated often enough, is to encourage the student to explore the inferential relations between what has just been learned and what he already believes to be the case. In this way, the new information is, as it were, woven into an existing web of belief. Less metaphorically, we might say that, as a result of such questioning on the part of the teacher, the student integrates the new information with existing beliefs and intuitions. And it is quite plausible to suppose that this process of integration has the effect of fixing, or, as Socrates might put it, "tethering", the new information in the mind of the student.

As well as being intrinsically plausible, the above interpretation of Socrates' remarks about the transition from true belief to *episteme* explains some otherwise curious features of the dialogue. In particular, it makes sense of the argumentative strategy that Socrates adopts throughout the *Meno*.

Recall from chapter 1 that, before Meno raises the paradox about the possibility of rational inquiry, Socrates tries to get him to come up with a satisfactory definition of virtue[23]. After a few false starts, Meno makes two suggestions: (i) virtue is the capacity to govern men[24]; and (ii) virtue is the desire for fine things, and the ability to acquire them[25]. Neither effort succeeds, as Socrates demonstrates; but the interesting thing, from our point of view, is *how* he does this[26].

In response to the first suggestion, Socrates asks Meno whether he thinks it is virtuous for a slave to govern his master. Meno concedes that it is not, thereby acknowledging that his proposed definition of virtue is flawed. This is an example of a quite general dialectical strategy of Socrates, used throughout the *Meno*, and in other dialogues. The strategy works by showing that views professed by his opponents conflict with other beliefs that they are unwilling to give up. In this case, Meno believes the both of the following:

1. Virtue is the ability to govern men
2. It would not be virtuous of a slave to govern his master.

He believes these propositions in the sense that he would answer "yes" if asked, "Is virtue the ability to govern men?", and "no" if asked, "Would it be virtuous of a slave to govern his master?"[27]. However, what he has evidently not noticed before - what it takes Socrates to bring to his attention - is that these two answers are incompatible. Meno has access to all of the information he needs to criticise his first attempt to define virtue, but he is unable or unwilling to use it until he is forced to by persistent interrogation. At this point, the contradiction that Socrates exposes is so glaring that it can no longer be ignored. In short, Socrates shows, through a carefully chosen series of questions, that the definition of virtue that Meno proposes cannot be integrated in a stable fashion with his other beliefs.

[23] See chapter 1, pp9-10 for further discussion of this part of the *Meno* dialogue.
[24] *Meno*, 73d.
[25] *Meno* 77b.
[26] For brevity's sake I deal only with the first definition in what follows; similar remarks apply to the other.
[27] Something like this sense of belief is explored by L. Powers in "Knowledge by Deduction", although he chooses to think of it as a weak variety of knowledge.

There is an interesting and important contrast between the previous example involving the two lowest prime numbers, and this one. In the former case, the mark of superior understanding in the student was the ability to answer questions to which he had not explicitly been given the correct answer. In the latter case, superior understanding is exhibited in a *reluctance* to answer certain questions: Meno's readiness to offer pat definitions of virtue just shows that he does not know what he is talking about. So, we might say, *thinking things through sometimes forces us to recognise that we do not understand something as well as we thought we did.* An account of understanding must be able to cope with both kinds of case.

To sum up the argument of this section:

1. At the end of the slave boy dialogue, Nemo has knowledge, but does not, claims Socrates, have *episteme*. This implies that "knowledge" is not the best translation for *episteme*, the goal of inquiry.

2. "Understanding" is a better translation, since, like *episteme*, it comes in degrees, and cannot be equated with having any old set of justified true beliefs.

3. Justified true beliefs may be converted to understanding[28] by asking questions that invite the student to explore the inferential relations between those beliefs and other beliefs they hold.

4. As a result of such questioning, the student may come to integrate the new information into her existing web of belief.

5. The end result of this process is a state in which one's beliefs about a given domain enjoy a distinctive and valuable form of stability; this stability is characteristic of understanding.

Of the above, points 3 and 4 relate to what, in section 3.2, I called *dynamic* understanding. Having discussed point 3 at some length in this section, I propose to examine point 4 in the

[28] I shall assume that the second point is correct, and use "understanding" in place of *episteme* from now on.

next. Point 5 relates to the static aspect of understanding, and is discussed in detail later in the thesis, beginning in section 3.5.

3.4 Minimal Comprehension

Towards the end of the previous section, I suggested that new information is understood to the extent that it is integrated with the rest of one's beliefs, and that this process of integration corresponds with what I called dynamic understanding in section 3.2. In this section I aim to be more precise about what is involved in an episode of dynamic understanding. To do this, I introduce the notion of what I call *minimal comprehension*. Roughly speaking, this is the least understanding of new information that someone can have, consistent with their having registered the information at all. An episode of dynamic understanding occurs whenever someone gets beyond this minimum level of comprehension, and comes to be able to solve new problems, or to recognise their ignorance, as a result[29].

Let's begin by introducing some terminology and notation that will simplify the exposition[30]:

Let the individual that we are interested in be S. As we are concerned solely with non-linguistic understanding in this thesis, we can assume, for the sake of simplicity, that S is a competent speaker of the language[31].

Let us suppose that S *is given a new piece of information,* P. For the sake of convenience of exposition, assume that this information takes the form of a single proposition. Ultimately, I believe, nothing hinges on this: P could just as easily represent the content of a lecture, or the information contained in an episode of visual perception, say[32].

[29] Or even, as we shall see, to reject the new information on the grounds that it conflicts with their existing beliefs.

[30] Although I use symbolic notation in what follows, this should not be interpreted as an attempt to introduce a formal calculus for representing episodes of dynamic understanding; it is, primarily, a notational convenience.

[31] It may be objected that coming to understand a new discipline often involves learning the meaning of many novel theoretical terms, so linguistic and non-linguistic understanding cannot be separated in this way. Although I am sympathetic to this point of view, I do not believe we should allow it to unduly complicate our account of dynamic understanding. Hence the assumption that S merely has competence in the language rather than complete mastery of it.

[32] Though, as discussed later in the section, there are a few subtleties to be aware of in these cases; see pp105-106.

Let Q *be the set of questions*[33] *that* S *understands*[34], *given her level of linguistic competence.* I assume, for the sake of simplicity, that acquiring the new information, P, has no effect upon this set, this being consistent with our confining our interest to cases of non-linguistic understanding.

Finally, let R_0 *be the set of responses* S *would make to members of* Q, *prior to acquiring the information* P, *and* R_1 *be the set of responses* S *would make to members of* Q, *after acquiring the information* P.

This last item, the notion of a set of responses to the members of Q, requires further comment. Firstly, the only responses we are interested in are those that S would give *under normal conditions*. Hence the responses that S makes when drunk, or otherwise disorientated, do not count[35]. Secondly, there is no requirement that S be able to respond *immediately* to all members of Q; it is quite all right for S to take time out to perform a few calculations, or follow through some complex chain of reasoning, if that is what she thinks the question calls for. In short, we are interested in S' *considered* responses to the members of Q. Thirdly, by a response, I mean a reaction to the question *qua that particular question.* This means that, among other things, the following do not count as responses, in this sense:

- Physical reactions, such as punching the questioner, or storming out of the room[36].
- Such meta-questions as, "Why do you ask?", that are concerned not so much with clarifying the question being asked, as inquiring into the motives behind it.
- Answers to questions other than the one asked, as sometimes occurs when politicians are interviewed.

[33] Framing the analysis in terms of questions rather than problems (say) ties it more closely to the language of S than is strictly necessary, but I don't think that this is a big problem. At any rate, the strategy is a natural one, given that the present line of inquiry was inspired by a Socratic dialogue.

[34] "Understands" here is used in a minimal sense, so there is no circularity in the account of understanding. S understands a question in this thin sense as long as she is familiar with the terms it employs and their syntax. Alternatively, we could let Q be the set of *all* questions, insofar as that idea makes sense, and allow "I don't understand the question" to be an acceptable response. That way, there's some hope of ending up with an account of minimal comprehension that applies to the special case of linguistic understanding.

[35] I accept that what counts as normal is hard to pin down; but I think that there are enough situations that we regard as definitely *not* normal to make the idea of normal conditions serviceable.

[36] Though a nod or shake of the head could count as a response to a question, since these gestures have conventional meanings.

Whilst the following, among other things, *do* count as responses in the intended sense:

- *Answers to the question*. Of course, the answer need not be correct, or even justified by the evidence at S' disposal; for our purposes, all that matters is that the answer be *sincere*, in the sense that it is an accurate reflection of the cognitive state of S. Hence Meno's claim that virtue is the capacity to govern men is *his* answer to the question, "What is virtue?" in this sense, albeit an incorrect one for which he has no justification whatsoever[37].

- *Professions of ignorance*, as might be expressed by saying, "I don't know", or words to that effect. As we saw in our discussion of the *Meno*, Socrates often takes this path.

- *Rejection of the question*. For an example of this type of response, consider the question, "Who is the present king of France?". Someone who does not know that France is a republic might react to this question in one of the first two ways, most likely the second. However, the correct response is to point out that the question incorporates a false *presupposition* - that France has a king. This counts as a rejection of the question, rather than an admission of ignorance, because, in this case, there is nothing to be ignorant of[38].

- *Meta-questions aimed at clarifying the nature of the original question*. For example, if someone asks me, "How far is that planet from the Sun?", I may respond by asking, "Which planet do you mean?". This is a genuine response to the original question, since its purpose is to determine precisely what is being asked, with a view to then answering the question.

The final point about the response sets R_0 and R_1 is rather more subtle. Sometimes, the fact that S has just given a certain response to one question may mean that S gives a response to a second question that differs from the one that S would have given had the first question not

[37] Hence, the sense in which I am using "answer" differs from the one found in texts on erotetic logic. My usage maps neatly onto what I referred to in chapter 1 as the "say what you think" constraint; see p21.

[38] The example given makes what is known as an *existential* presupposition that happens to be false. There are also *attributive* presuppositions. For example, "Does the Arts Tower like sport?" presupposes that the Arts Tower is the kind of entity to which the predicate "... likes sport" may be applied. For more on presuppositions of both kinds, see the early pages of S. Bromberger, "What We Don't Know When We Don't Know Why".

recently been asked. This is what happens in the *Meno* dialogue, for example. At first Meno responds to the question, "What is virtue?" confidently, with the claim that it is the capacity to govern men. But after he concedes that it is not virtuous for a slave to govern its master, he is reluctant to repeat this answer. So, his response to the question "What is virtue?" changes as a result of answering the further question, "Is it virtuous for a slave to govern its master?". To avoid the difficulty that this example highlights, we must imagine the members of Q being put to S one at a time, "resetting" S' cognitive state after each one. In other words, R_0 and R_1 represent sets of responses that S would give to the members of Q, if each one was the *only* question asked.

Bearing in mind all of the above, let us ask ourselves what is the very *worst* that S can do in terms of comprehending the new information P. One possibility is the following:

M1: S minimally comprehends P *iff* $R_1 = R_0$

This states that, S responds to the members of Q, after acquiring the new information, P, in exactly the same way that she would have done prior to acquiring that information. In other words the new information does not make any difference to the responses given by S: even the question, "Is it the case that P?", meets with the same response - an admission of ignorance say - than it would have done had S never been informed of P. Put this way, the situation described by M1 does appear to describe a genuine worst case; but it also begins to look more like a bizarre failure of memory than a failure of understanding*per se*. So, at first sight, M1 describes a case that seems too outlandish to be of much use to us. Later, we shall see that this is not so: under certain circumstances, an individual that satisfies M1 has demonstrated non-minimal comprehension! For the time being though, we ought to modify our definition of minimal comprehension to look something like this:

M1*: S minimally comprehends P *iff* $R_1 = R_0 + R_P$

Where R_P represents the set of responses given by S to members of Q_P, a small subset of Q, consisting of just those questions that *could* be answered - answered *correctly*, that is - given only the information P, and competence in the language[39].

[39] If P represents the informational content of a book, say, rather than a single proposition, Q_P may, in fact, be quite a large set. I look briefly at cases like this, towards the end of this section.

For example, suppose that P represents the information that Paris is the capital of France, and that this is new information, as far as S is concerned, although she does know that the Eiffel Tower is in Paris. Then Q_P includes such questions as, “Is Paris the capital of France?”, and, “What is the capital of France?”, and R_P represents the fact that S can now answer these questions. However, Q_P does *not* include questions such as, “Is it true that the Eiffel Tower is in the capital of France?”, since this cannot be answered given only the information that P. A second example highlights the role played by the requirement that S be a competent speaker of the language: if P is the information that Cliff Richard is a bachelor, Q_P includes the questions, “Is Cliff Richard male?”, and, “Is Cliff Richard married?”, since S is supposed to be aware that a bachelor is an unmarried man[40].

But this revised definition of minimal comprehension is still not quite correct. To see this, consider that, in the illustrative examples of the previous paragraph, S would have had*some* way of responding to the members of Q_P, prior to acquiring the new piece of information - most likely with admissions of ignorance. So, R_P does not represent a small set of *new* responses, distinct from, and in addition to, the responses given by S, prior to acquiring the information P; rather, R_P represents the fact that a small subset of the responses originally given by P has *changed* as a result of the availability of new information. What this means is that the addition sign in $M1^*$ is misleading; something like this is what we’re really after:

$M1^{**}$: S minimally comprehends P *iff* $R_1 = \Delta_P(R_0)$

Where Δ_P is the function that replaces *all and only* the responses that S gives to members of Q_P prior to acquiring the new information with those that S gives afterwards[41].

The observant reader may notice that $M1^{**}$ includes our original definition, M1, as a special case (when *none* of S’ responses to the members of Q_P change after S acquires the information that P). This implies that our definition allows, for the time being, at least, degrees of minimal comprehension, ranging from the bewildering case described by M1, to the “best” case in which S is able to answer *all* members of Q_P correctly as a result of

[40] This presupposes a relatively sharp analytic/synthetic distinction, of course, but I believe that this is an acceptable idealisation for the purposes of developing a working model of dynamic understanding. If, in the real world, the boundary between the analytic and the synthetic is often blurred, this just implies that the concept of minimal comprehension is not sharply defined.

[41] Note that Δ_P may leave some, or even all, of S’ responses to the members of Q_P unchanged (replacing them, as it were, with themselves). Later we shall need to amend to this specification of Δ_P, to deal with an objection to $M1^{**}$ (the objection referred to in the last sentence of the next paragraph).

acquiring the information P, S previously having been ignorant of the answers[42]. It turns out that one of the objections to this definition, to be described shortly, will force us to revise it to rule out the possibility of a range of states of minimal comprehension[43].

Given $M1^{**}$, we can simply define *non*-minimal comprehension as follows:

N1: S comprehends P to a greater than minimal degree *iff* $R_1 \neq \Delta_P (R_0)$

And say that an episode of dynamic understanding occurs whenever S comprehends a new piece of information, P, to a greater than minimal degree.

The above definitions reflect the idea, arising from our discussion of the*Meno*, that dynamic understanding is a matter of integrating new information with what one already believes to be the case. Where there is only minimal comprehension, there is little integration, and this is reflected in the minimal difference between R_1 and R_0. As the degree of comprehension of the new information increases, however, so does the gap between R_0 and R_1[44]. At the far end of the spectrum, we find such remarkable feats of comprehension as Einstein's formulation of the Special Theory of Relativity, as triggered, ultimately, by his determination to take the negative result of the Michelson-Morley experiment at face value.

Defining minimal and non-minimal comprehension in terms of the difference between R_1 and R_0 also captures some of the richness of the concept of understanding, since the two sets may include responses of several different types. This point was glossed over at the end of the previous section, where we discussed just two examples of the difference that thinking things through[45] makes: Nemo answering a question about prime numbers that no one had told him the answer to; and Meno, with a little help from Socrates, recognising his ignorance about the nature of virtue. In the former, ignorance gives way to firm belief, whilst, in the

[42] Of course, if we're concerned about how well S has understood the new information, even minimal comprehension, at its best, will not impress us much.

[43] See p100ff for more details.

[44] It might be feasible, in principle, though probably difficult in practice, to specify an, at least partial ordering of the states of non-minimal comprehension. I do not attempt this here, as the main purpose of the section is simply to say what dynamic understanding is.

[45] The phrase "thinking things through" is deliberately vague. In the examples given, it is a matter of putting new information together with one's existing beliefs, and drawing valid *deductive* inferences. But it should also, as a bare minimum, include making appropriate *inductive* and *abductive* inferences. In addition, I think there is a good case for including reasoning based on analogy, insofar as it is not reducible to a combination of the other three kinds of inference. In this respect, I agree

latter, firm belief gives way to the acknowledgement of ignorance. Thanks to the typology developed earlier in this section, we can now envisage further possibilities: acquiring new information may lead us to reject questions we previously accepted as legitimate[46], or it may force us to accept that a previously rejected question is legitimate after all, and that we are simply ignorant of the answer. It is a virtue of M1** and N1 that they are broad enough to embrace all of these possibilities.

Although our definitions elegantly capture many of our intuitions about dynamic understanding, a number of objections to them indicate that some refinement is necessary. Firstly, suppose that S comes, for whatever reason, to believe, that Tony Blair has a glass eye. As a result of acquiring this new information, P, S forms, seemingly at random, lots of other new beliefs - that T. S. Eliot wrote *Four Quartets*, John Suckling invented the game of cribbage, and so forth. In this example, N1 is satisfied, since:

1. Because of her newly, if deviantly, formed beliefs, S now *answers*[47] the question, "Who invented the game of cribbage?", instead of admitting her ignorance.

2. The question, "Who invented the game of cribbage?" is not a member of Q_P. That is, it is not a question that can be answered correctly if all the information we have is that Tony Blair has a glass eye.

3. So, the differences between R_1 and R_0 cannot be accounted for solely in terms of differences in the responses given by S to members of Q_P.

Therefore, S has non-minimal comprehension of P. But this change in the way that S responds to questions that have nothing to do with the new information about Tony Blair strikes us as an aberration, rather than as a manifestation of understanding[48]. A similar objection applies to our definition of minimal comprehension, since M1** places no constraints on the way in which S' responses to the members of Q_P may change in the light

with Neil Cooper's reservations about restricting understanding to the ability to draw appropriate deductive inferences, as expressed in "The Epistemology of Understanding" (especially section III).

[46] This often happens in the history of science, the classic example being the replacement of Aristotelian mechanics with the Newtonian paradigm.

[47] Correctly answers, as it happens, though this is not important, in this case.

[48] Unless, that is, S previously believed that *if* Tony Blair has a glass eye, *then* Eliot wrote *Four Quartets*.

of the new information, P[49]. Hence, to return to our example, there is nothing to stop S from responding to the question, "What is Tony Blair's eye made out of?" (a member of Q_P, let us say) with the answer, "Camembert", should the mood take her. But this is clearly absurd, and does not deserve the title *comprehension*, minimal or non-minimal.

Roughly speaking, the problem in both cases is that there is no logical connection between the newly acquired information, and the further conclusions that S reaches. This suggests that we need to introduce a *normative* element into our definitions of minimal and non-minimal comprehension[50]: S should be able to account for whatever discrepancies there are between R_1 and R_0. In short, we need to add something like the following condition to M1** and N1:

A1: S sees P as *warranting* the discrepancies between R_1 and R_0

However, this is still not enough, for S may subscribe to deviant rules of inference. If, for example, S thinks that it is always permissible to infer a proposition about cribbage from a proposition about glass eyes, then she will be able to account for her new belief about the game's inventor in the first of the above examples. And, again, a similar objection applies to the definition of minimal comprehension. Hence, a third and final condition is necessary:

A2: P *does* warrant the discrepancies between R_1 and R_0

Of course, A2 assumes that we have a satisfactory general account of when inferences *are* warranted. This may be the case as far as relatively simple systems of deductive logic are concerned, but standards of rationality for inductive and abductive inferences are notoriously elusive. However, this is a separate philosophical project; the account of understanding I am defending can simply plug in its results, as and when they arrive[51].

A second objection to our definition of non-minimal comprehension is that it suffers from *conjunctivitis*. To see what this means, suppose that one of the things that S believes, before acquiring the information that P, is P*. Suppose also that, before learning that P, S responds to the question "Is it the case that P *and* P*?" with an admission of ignorance. Note that this

[49] Though, later, this will change. See my discussion of the third objection to our definitions of minimal and non-minimal comprehension on p100ff, for further details.

[50] But note that this normative element is later removed from the definition of *minimal* comprehension. See my discussion of the third objection, on p100ff, to find out why.

question is not a member of Q_P, since S will not be able to answer it correctly, given only P, the information P^* being also required[52]. After learning that P, however, S may change her response to the question, "Is it the case that P and P^*?", since she is now in a position to answer it in the affirmative. This case satisfies our definition of non-minimal comprehension, for:

1. There is at least one question that is not a member of Q_P, to which the response given by S changes, as a result of acquiring the new information, P.

2. S, we may suppose, sees the change in her response to this question as *warranted* by the new information.

3. The new information P, *does* warrant the change in S' response to this question.

The problem is that this conclusion seems to justify calling *any* string of random conjuncts an exhibition of understanding, yet this cannot be correct. To see the absurdity of this conclusion, just substitute, "Peter Parker is Spiderman", and, "Donald Fagen is a member of Steely Dan" - or any other pair of unrelated propositions - for P and P^*, in the above example. Clearly, the results are not the sort of thing we would wish to exhibit as episodes of dynamic understanding.

The conjunctivitis problem is especially vexing, because our definition of non-minimal comprehension does not suffer from the analogous problem of *disjunctivitis*; "Is it the case that P *or* P^*" is a question that *can* be answered correctly, given only P, regardless of whether S previously believes that P^*, and so qualifies for membership of Q_P. Hence forming random disjunctive propositions which include the new item of information can never demonstrate more than minimal comprehension of it. This shows that our definitions are asymmetrical in their treatment of conjunction and disjunction: the former can be a manifestation of non-minimal comprehension, whereas the latter can never be a manifestation of anything more than minimal comprehension. Yet, intuitively, we would expect the two cases to be classified in the same way, neither getting beyond minimal comprehension; it is hard to see how this preferential treatment of conjunction can be justified.

[51] Or, at least, I see no reason why it should not be possible to do this.

There are several ways of responding to this worry. One option is to add another condition to the definition of non-minimal comprehension designed specifically to rule out conjunctivitis. This has the advantage of securing symmetrical treatment for disjunction and conjunction, neither now counting as true exhibitions of dynamic understanding, but has the drawback of being blatantly *ad hoc*. A second possibility is to bite the bullet by conceding that the ability to form arbitrary conjunctions *is* a manifestation of understanding, albeit not a very impressive one. There are three problems with this suggestion. Firstly, it may be necessary, but difficult, to say more about *why* this way of combining the new information, P, with one's existing beliefs is so unimposing, as feats of comprehension go. Secondly, the concept of minimal comprehension becomes a *very* thin one if this proposal is adopted - so thin that conjunctivitis counts as a step beyond it - and loses much of its philosophical interest in the process. And finally, it leaves us with that nagging, apparently inexplicable, asymmetry in the treatment of disjunction and conjunction.

Perhaps the best strategy is to address this issue head on, by asking what it is about arbitrary disjunctions and conjunctions that makes them seem suspect in the first place. One idea is that neither of them generate genuine beliefs. Upon being informed that P, we may then answer the question, "Is it the case that P or P*?", in the affirmative, but this is just an expression of our belief in P (or P*) not the expression of an independent belief in the disjunctive proposition. Likewise, having a belief that P *and* a belief that P* may incline us to answer the question, "Is it the case that P and P*?", in the affirmative, but there is no independent belief in the conjunctive proposition. This may sound controversial, but it accords with the plausible view that an essential feature of beliefs is that they have a role in the explanation of action and inference. If this is correct, then we should, perhaps, ascribe to an individual no more beliefs than are warranted by this explanatory need[53]. From this point of view, the ascription of disjunctive and conjunctive beliefs in the two cases just mentioned is simply otiose[54].

[52] Unless P* is a logical consequence of P, which, for the purposes of our example, we may assume is not the case.

[53] These remarks are inspired by T. Grimaltos and C. Hookway, "When Deduction Leads to Belief", especially, sections 2 and 3. They develop this line of thought partly as a response to the examples used by Gettier in "Is Justified True Belief Knowledge". However, it is natural to insist that, in the analysis of the concept of knowledge, "belief" is used in a much thinner sense than Grimaltos and Hookway suggest - S believes that p when they are prepared to assent to it, perhaps. If this is correct, then the psychologically thick sense of "belief" that they talk about is irrelevant. Grimaltos and Hookway spend a considerable amount of time replying to this point, but we don't have to: our concern all along has been the psychologically real beliefs that guide inquiry.

[54] Some problems remain though. Suppose, for example, that S believes both that Q, and that *if* P & Q, *then* R. S then learns that P, and from the conjunction P & Q, deduces R. Here, S' belief in the

As far as the discussion of minimal and non-minimal comprehension goes, the above line of thought suggests that we need to be more restrictive about what should count as a response to a member of Q, the set of questions that S understands, given her level of linguistic competence. We decided earlier that we are only interested in considered responses, given by S, under normal conditions. To these two conditions, we should now add that the only responses we are interested in are those that express S' genuine beliefs, in the psychologically thick sense, outlined above.

The third, and final, objection to the definitions of minimal and non-minimal comprehension that I wish to consider is raised by the following example. Suppose that, before being informed that P, S believes a number of propositions that, together, entail not-P; call this set of propositions, A. For the sake of the example, assume that the chain of reasoning that leads from the members of A to the conclusion not-P is long and complex. S rejects the information, P, when it is offered, on the basis of this chain of reasoning. In this case, we are strongly inclined to say that S has exhibited non-minimal comprehension, having rejected the information that P for a good reason, namely, that it does not fit in with a number of her previous beliefs. This suggests that seeing that a piece of new information cannot be integrated with one's existing beliefs[55] may also qualify as an episode of dynamic understanding.

The problem is that this possibility is not catered for by our definition of non-minimal comprehension. As things stand, S does not get beyond minimal comprehension, in the above example, since being informed that P has not changed any of S' previous beliefs; in particular, it has not altered S' belief that P is false, although it has made this belief more prominent in her mind. Since none of S' beliefs have changed, neither should the responses, given by S, to the members of Q. Hence, in this case, $R_1 = R_0$, and this, as I pointed out earlier[56], describes a special case of the current definition of minimal comprehension.

The key to resolving this difficulty is to recognise that there are really *two* things going on in the above example:

conjunction *does* play a role in the explanation of the inference that she makes. Analogous cases may be constructed with disjunctive propositions. However, our original worry was that the deduction of a trivial conjunction or disjunction would count as an exhibition of understanding according to our definitions; and, in the example just described, S has already moved beyond this.

[55] Not without sacrificing more of them than one is willing to, at any rate. See below, for more on this point.

[56] See p94.

1. S infers not-P from the members of A.
2. S recognises that this conclusion is inconsistent with P, and decides to reject the new information.

The example is deceptive, because it presents itself as a counterexample to our definition of non-minimal comprehension, in virtue of the chain of inference S follows to arrive at the conclusion not-P. But, if we think of non-minimal comprehension as always being a matter of *comprehension of P with respect to one's beliefs prior to being informed that P*, the fact that S follows this chain of inference is simply irrelevant. As stated at the beginning of this section, dynamic understanding is a matter of *integrating* new information with one's existing beliefs; so reasoning - however impressive - that relies *only* upon one's existing beliefs, or *only* upon the new information, just drops out of the picture. Hence, inferring the consequences of one's current beliefs does not count as an exhibition of non-minimal comprehension; in particular, S' inferring not-P from the members of A is not an example of non-minimal comprehension *of* P *with respect to* S' *previous beliefs*. This is not as counterintuitive as it might sound, for we can still admit that this inference demonstrates non-minimal comprehension, on the part of S, of the last member of A she accepted, say, with respect to the beliefs she had at that time.

So, once we specify clearly what our definition of non-minimal comprehension is a definition *of*, the alleged counterexample to it, described above, appears less threatening. But what about the second aspect of the example, S' decision to reject the new information, P, on the grounds that it conflicts with her existing beliefs? It turns out that this *does* have a bearing upon our definition of non-minimal comprehension. To see this, consider a simple example, where the contradiction between my existing beliefs, and the item of new information is quite blatant.

Suppose that I read in the *News of the World* that Elvis Presley is alive, and living in Ilford. This conflicts with my memory of his death being reported in the *Guardian* and on the TV news several years ago. Of course, I shall reject the more recent report as probably false, on the basis of my existing beliefs; but why do I take the *Guardian*'s word for it, rather than the *News of the World*'s? Because, for one thing, I believe it is a more reliable source of accurate information than it's tabloid rival - which isn't to say, of course, that I believe it is infallible. Also, in this case, the intrinsic implausibility of the idea that Elvis could have faked his own death[57],and lived undetected for many years, given the intense public interest in his life, counts heavily against the *News of the World*'s report. In other words, there is a chain of reasoning behind my decision to reject the *News of the World's* report, that looks something like this:

1. The *News of the World* report contradicts my belief that Elvis Presley is dead, so I must reject either the report, or my existing belief.

2. The *News of the World* is, generally, a less reliable source of information than the source of my existing belief. In addition, the report is intrinsically implausible, in the light of my other beliefs.

3. *Therefore* the *News of the World* report is probably false, and I should continue to believe that Elvis is dead.

Although our original example is more complex than this one, in that a chain of reasoning is required to highlight the contradiction, there is no reason to suppose that it does not resemble this example in other respects. In particular, the lesson of the above example - that inconsistencies in one's belief set may be resolved by evaluating the reliability of the sources of those beliefs, and their plausibility in the light of one's other beliefs - seems to be a general one, equally applicable to the original example. For in that example, S must still find some way of resolving the inconsistency: should she reject P, or should she reject one or more of the members of A, the set of beliefs that, together, entail not-P? It is difficult to see how S can make this decision - make it *rationally*, at least - without following something like the chain of thought that leads me to reject the *News of the World* report, in the second example. The objection to our definition of non-minimal comprehension glosses over this point, by presenting the rejection of the new information as a *fait accompli*.

But the important thing, from our point of view, is that the type of reasoning described in the preceding paragraphs *should* count as a manifestation of dynamic understanding. Even if the new information, P, is ultimately rejected, as a result of this process, S has still made an effort to integrate P with the rest of her beliefs, and has only rejected P because this cannot be done without sacrificing other beliefs that are more tenable. Intuitively speaking, in this example, S exhibits the sort of reflective attitude to new information that we think of as

[57] Or could have risen from the dead!

characteristic of understanding. The problem is that, as we noted earlier, the current definitions of minimal and non-minimal comprehension would classify this case as the former, rather than the latter. So, as it turns out, the original example*is* a counterexample to our definitions, but not for the reason we thought it was!

The way to deal with this problem is to recognise that our*default* attitude to new information is to accept it as *true*. The principle underlying this attitude is that most of what we read or hear about is true; call this the *trust principle*. To the extent that we ever think about this principle, we probably see it as warranted by the general trustworthiness and reliability of the people we have encountered. But it seems unlikely that the trust principle could ever be definitively justified in this way; the ultimate justification for it is likely to be a pragmatic one, adoption of the principle being seen as preferable to the kind of epistemic paranoia that would appear to be the alternative.

The examples discussed in the preceding paragraphs show that, although trust is, in general, a good thing, it is not an attitude to be adopted*blindly*, in all circumstances. Sometimes, new information must be regarded with suspicion, one such occasion being when it conflicts with what we already take to be the case. In such circumstances, as argued above, one must think about the reliability of the sources of one's information, and the plausibility of one's beliefs, in the light of each other. This suggests that someone who demonstrates no understanding of new information may be suffering from a form of*gullibility*. Instead of adopting the critical, reflective attitude to new information, that is occasionally required, such an individual always adopts the default position of taking new information on trust, as the truth. Although this is often the rational response, sometimes it is an inadequate one; sometimes a little more *understanding* of what one has just been informed of is called for.

The above observations supply us with enough insight into the nature of our problem to attempt a new formulation of our definitions. As a matter of fact, the definitions can stay pretty much as they are - aside from making it explicit that they refer to comprehension of P with respect to S' beliefs prior to being informed that P. What needs to change is the description of the function Δ_P that appears in them. If our default attitude to new information is to accept it as the truth, and if this is the attitude is adopted unthinkingly when someone exhibits minimal comprehension, then Δ_P is *not* the function that replaces all and only the responses given by S to the members of Q_P before acquiring the information that P, with those given afterwards. Rather Δ_P is the function that replaces S' responses to the members of Q_P before being informed that P, with the *correct* answers to the members of Q_P, *on the assumption that P is true*[58].

The normative conditions appended to $M1^{**}$ and N1, also need to be reconsidered. As far as the definition of minimal comprehension goes, a normative element is no longer required. For minimal comprehension now stands revealed as an essentially mechanical process, in which S unthinkingly adopts the default attitude towards the new information P, accepting it, and all of its logical consequences as the truth, regardless of whether it conflicts with her existing beliefs. In contrast, the definition of non-minimal comprehension still requires a normative element. The difference is that the new specification of Δ_P means that, in those cases where S rejects the new information, P, she should be able to justify this decision. Of course, as before, the justification that S gives should be a sound one.

To summarise this discussion of minimal and non-minimal comprehension:

S *minimally comprehends* new information, P, with respect to her existing beliefs, *iff* $R_1 = \Delta_P (R_0)$.

S has *non-minimal comprehension* of new information, P, with respect to her existing beliefs, *iff*:

N1: $R_1 \neq \Delta_P (R_0)$
A1: S sees P as *warranting* the discrepancies between R_1 and R_0
A2: P *does* warrant the discrepancies between R_1 and R_0

Where: R_0 and R_1 are sets of responses given by S to members of Q, the set of all the questions she understands[59], before and after being informed of P; Δ_P is the function that replaces all and only S' responses to members of Q_P before being informed of P, with the *correct* answers to the members of Q_P, *on the assumption that P is true*; and Q_P is the set of all questions that can correctly be answered, given *only* the information P, and competence in the language.

[58] The last clause is required since, if P is false, the correct answers to some of the members of Q_P will find their way into R_1 without any, as it were, positive effort to comprehend P's relationship with her other beliefs, on the part of S.

[59] Where this refers to *linguistic* understanding, as opposed to non-linguistic; see section 3.2, for further details.

Dynamic understanding is exhibited whenever an episode of non-minimal comprehension takes place.

Before leaving the topic of minimal and non-minimal comprehension, let's look briefly at the case where P is more than just a single item of propositional information[60], as might be the case if P represents the content of a philosophy lecture, or a chapter of an atomic physics text, say. The first thing to note about such cases is that Q_P, the set of questions that can be answered correctly, given only the information, P, is much larger than its counterpart when P is just a single proposition. At first sight, this fact, together with the latest revisions to the definitions of minimal and non-minimal comprehension may appear to present a problem. Remember that both definitions now include the function, Δ_P, which substitutes the *correct* answers, on the assumption that P is true, to the members of Q_P, for the responses given by S, before being informed of P. The problem is that expecting S to have the answers to *all* of the members of Q_P, in cases where P represents a large amount of new information, seems highly unrealistic. Think, for example, of the number of questions about the contents of an advanced atomic physics text that could legitimately be asked; how can we reasonably expect S to be able to generate all of those answers for herself? Moreover, even if S *is* able to answer all of these questions, our definitions do not guarantee that, in doing so, she has achieved more than minimal comprehension of the text – can this be fair?

Yes it can. To see this, we just need to remind ourselves that our definitions are definitions of minimal or non-minimal comprehension *of P, with respect to one's existing beliefs*, dynamic understanding being exhibited when new information is *integrated* with one's other beliefs. So, as stated previously, any reasoning by S that relies *only* upon the new information, however much new information is involved, and however impressive the reasoning, is irrelevant as far as the definitions of minimal and non-minimal comprehension are concerned. This means that, for the purpose of working out the definitions, we can make as many assumptions and idealisations as we like, regarding S' reasoning capacities, as long as the reasoning in question does not, as it were, straddle the boundary between the new information and the old. So, as far as our definitions of minimal and non-minimal comprehension are concerned, we can just stipulate that S *does* know the answers - on the

[60] As a matter of fact, we have already encountered such cases – in our discussion of the last objection, above. The resolution of that difficulty depended, partly, upon recognising that we resolve inconsistencies in our belief sets by evaluating the reliability of the sources of the beliefs in question. This implies that information never comes in discrete packets, as we have assumed so far; even in the simple case discussed earlier, I register not only the information that Elvis is alive (allegedly), but also the associated information that Elvis is alive *according to the News of the World*.

assumption that P is true - to all of the members of Q_P. We can even imagine *giving* all of these answers to S, since we are interested in what S does with them afterwards, and, in particular, in how S relates the new information with what she already believes to be the case.

Of course, S has the opportunity to manifest dynamic understanding in the course of reading the advanced atomic physics text. But it should now be obvious that this possibility is already catered for by our definitions. In effect, we just need to think of S as taking in the contents of the textbook one proposition at a time, relating each new item of information with what she has already learned, as she goes along. Or, if this seems like too much of an idealisation, think of S as pausing after each section or chapter of the book, and reviewing what she has just read in the light of her beliefs prior to reading it. It is a strength of our definitions that they cater for the possibility of a hierarchy of levels of understanding like this.

I conclude that the definitions of minimal and non-minimal comprehension above clarify our intuitions, such as they are, about the nature of dynamic understanding. They also enable us to make sense of a number of passages in the *Meno*, since it is clear from them that minimal comprehension is, crudely speaking, a matter of accepting everything we read or hear, without question, repeating it "parrot-fashion" as and when it seems appropriate. This, as the following episodes from Plato's dialogue show, is exactly what Meno does.

Episode One: Meno inaugurates the dialogue by asking Socrates whether virtue is something that can be taught. Socrates dodges this question by arguing that until one has determined *what* virtue is, one is in no position to state whether or not it can be taught. His admission that he is ignorant about the nature of virtue precipitates the following exchange [*Meno* 71c-d]:

M: What! Didn't you meet Gorgias when he was here?

S: Yes.

M: And you still didn't think he knew?

S: I'm a forgetful sort of person, and I can't say just now what I thought at the time. Probably he did know, and I expect that you know what he used to say about it. So remind me what it was, or tell me yourself if you will. No doubt you agree with him.

M: Yes I do[61].

Meno goes on to outline the account of virtue given to him by Gorgias, which Socrates easily demolishes.

Episode Two: As a warm-up for the main task of defining the nature of virtue, Socrates asks Meno to define the concept of shape. Meno asks Socrates to do it instead, and Socrates obliges with two definitions: (i) shape is the only thing which always accompanies colour; and (ii) shape is the limit of a solid. Meno then challenges Socrates to define the concept of colour:

S: Would you like an answer *a la* Gorgias, such as you would most readily follow?

M: Of course I should

S: You and he believe in Empedocles' theory of effluences, do you not?

M: Whole-heartedly.

[Socrates spells out the relevant details of the theory]

S: From these notions then "grasp what I would tell", as Pindar says. Colour is an effluence from shapes commensurate with sight and perceptible by it.

M: That seems to me an excellent answer.

S: No doubt it is the sort you are used to...[62]

Socrates goes on to slyly mock this answer, and to suggest that his definition of shape is much better than this definition of colour. These remarks appear to fall on deaf ears.

Similar episodes abound in the dialogue. Many of them seem insignificant, but taken together they form a vivid portrait of Meno as a man without a mind of his own, and who, consequently, is dependent upon other people for his opinions. Rather than say what he thinks, he repeats what Gorgias has told him, or recalls the words of a poet. The other side of Meno's reliance on other people for his opinions, is an aversion to any kind of mental exertion, as manifested in his reluctance to address the question of the true nature of virtue. In short, Meno has a problem, one which is memorably expressed by Jacob Klein, in his commentary on the dialogue. He notes that, whilst Meno can take in the opinions of other

[61] *Meno*, 71c-d.
[62] *Ibid.*, 76c-d.

people, and repeat them if appropriate, he is unable to examine those opinions critically. Klein concludes that Meno's "soul" lacks the features that make learning possible, for it is:

> ... indeed nothing but "memory", an isolated and autonomous memory, similar to a sheet, or to a scroll covered with innumerable and intermingled characters... it is a repository of opinions, but it cannot become a repository of knowledge[63].

To this damning diagnosis, we may now add that Meno seems incapable of exhibiting anything more than minimal comprehension with regard to the opinions of other people.

3.5 Models and Understanding

In this section, I introduce the concept of a model, and argue that it gives us another way of appreciating the nature of Meno's intellectual inadequacies, as discussed at the end of section 3.4. The claim is that, because Meno only ever responds to new information with minimal comprehension, he is not able to construct a model of the world in which he lives. This is a significant shortcoming because, as I argue in subsequent chapters, models provide us with a kind of intellectual orientation with respect to the domains they represent. This form of intellectual orientation is characteristic of what I called *static* understanding in section 3.2. This section therefore acts as a bridge between our discussion of understanding in the *Meno* (principally the dynamic variety), and the closing chapters of the thesis, which are concerned with the static aspect of understanding.

Let us begin our analysis of the concept of a model by considering an everyday example. On the desk, in front of me is a small-scale replica of Thunderbird 2, say; what is it about this object that makes it a model of the real thing? Sylvain Bromberger suggests[64] that it is in virtue of the fact that I can find out things about the real Thunderbird 2 by inspecting, measuring, and experimenting upon it. For example, I can determine how many wings Thunderbird 2 has by counting the wings on the object before me; or I can determine the colour of the real thing, by noting the colour of the object on my desk. This observation leads Bromberger to propose the following definition:

> M is a model of O, relative to a set of triples [Q_m, Q_o, T] iff in each triple, Q_m is a set of questions about M, Q_o is a set of questions about O, and T is an algorithm that translates

[63] Or understanding for that matter. The passage is from Jacob Klein, *A Commentary on Plato's Meno*, p136.

[64] In "Rational Ignorance".

any answer to a member of Q_m into an answer to a member of Q_o, and *correct* answers to the former into *correct* answers to the latter[65].

This is a useful starting point for our inquiry into the nature of a model, but a few alterations are required. Firstly, a minor problem with the above definition is that it presents the process of translation between model and object as one way only. The definition states that our answer to Q_m, may be converted into an answer to Q_o by means of the algorithm T, but does not tell us how to derive Q_m in the first place. But clearly, there must be some way of generating a member of Q_m, given the corresponding member of Q_o, if the model is going to be of any use to us. Hence, for the remainder of this section - and the thesis, for that matter - I shall use T to refer to both of the translation steps that are involved in the use of a model.

A second problem with Bromberger's definition is that its use of the term "algorithm" makes the process of translation between model and object sound too mechanical. Although moving from a proposition about the object to a proposition about the model, or *vice versa*, can be (and, perhaps, usually *is*) straightforward, there are occasions when it requires the exercise of ingenuity, of a kind not easily encapsulated in an algorithm[66]. It may be, for example, that the contour lines of a topographical map contain all the information necessary to answer the question, "Is there a hanging valley nearby?"; but it is not clear what *question* we should ask of the map in order to answer our question about the terrain it represents. For this reason, I think that it is best to think of T as a less well-defined type of transformation, one that may include an algorithmic component, but which could also incorporate heuristic elements. This leaves room for the possibility that it may require skill and creativity to come up with an adequate translation in any given case.

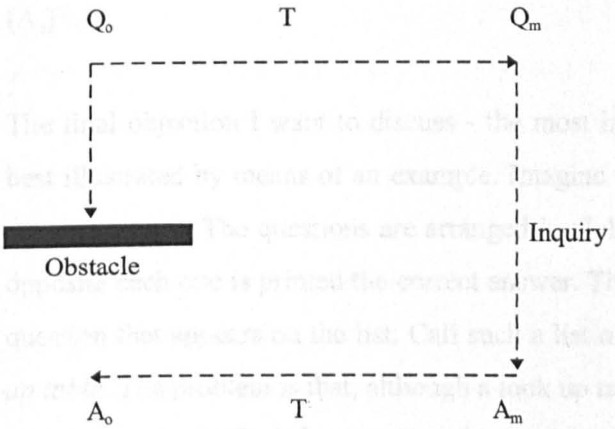
A third, rather more significant, objection to Bromberger's definition is that it makes the relation between a model and its object symmetrical. It is, for example, no less possible to answer questions about the model of Thunderbird 2 by inspecting the real thing, than it is to answer questions about Thunderbird 2 by examining the model. Hence, according to the above definition, the real Thunderbird 2 is a model of the model Thunderbird 2. As the awkwardness of the last sentence testifies, this conclusion is hardly in line with ordinary usage.

[65] *Ibid.*, pp139-40; I have modified Bromberger's definition very slightly.

[66] A related but distinct point: having determined the corresponding question about the model, it may require a good deal of ingenuity to *answer* it. Models are not intellectual panaceas.

The best response to this difficulty is to accept its conclusion, and concede that our use of the term "model" is a quasi-technical one. However, in taking this path, we should distinguish carefully between the semantic and pragmatic aspects of the modelling relation. *Semantic* questions concern the relationship between a model and the object it represents; it is this relationship that, as the present objection rightly points out, is symmetrical, according to Bromberger's definition. *Pragmatic* questions concern the way in which a model is *used*, in our inquiries, to answer questions about the object it represents, our main concern in this thesis. From a pragmatic point of view, the relationship between a model and the object it represents is *not* symmetrical, for *a model gives us an indirect way of answering questions about its object that is less costly than the direct approach*. The diagram below illustrates this point:

Figure 3.1: The Pragmatic Aspect of a Model



The diagram is reasonably self explanatory. It represents the process whereby a question, Q_o, about an object, O, is answered by being transformed into a question Q_m, about M, a model of O. The methods that would normally be employed to answer a question like Q_o, are not used, on this occasion, because they are judged to be too expensive[67]. The diagram represents the unavailability of this direct avenue of inquiry by means of the shaded bar marked *Obstacle*. The presence of an obstacle means that it is necessary to take an indirect route, *via* the model. This involves three stages:

[67] More than monetary costs may be involved, of course; typically, the amount of time and resources (raw materials, physical effort, intellectual effort, and so forth) a direct assault on the problem would consume, will also need to be brought into the reckoning.

1. Transform Q_o into Q_m, the corresponding question about the model.

2. Determine the answer, A_m, to Q_m, by whatever means of inquiry are appropriate to the model.

3. Transform A_m into A_o, the answer to the original question.

Returning to our simple example will make all of this plain. Suppose that I want to know the wingspan of Thunderbird 2 [Q_o]. Unfortunately, Thunderbird 2 is not around, and even if it were, it might be too difficult to measure its wingspan directly [Obstacle]. Luckily, there is a model of Thunderbird 2 in front of me. I can answer Q_o by answering the corresponding question about the model: what is *its* wingspan [Q_m]. This is easily determined with the aid of a ruler; suppose that the answer is 20cm [A_m]. Knowing that the scale of the model is 1:100 [T], I can convert A_m into an answer to Q_o: the wingspan of Thunderbird 2 is 20m [A_o][68].

The final objection I want to discuss - the most important one, from our point of view - is best illustrated by means of an example. Imagine that I have a piece of paper with a list of questions on it. The questions are arranged in alphabetical order, for ease of reference, and opposite each one is printed the correct answer. This arrangement enables me to answer any question that appears on the list. Call such a list of questions and associated answers a *look up table*. The problem is that, although a look up table is, intuitively, not a model, it seems to satisfy Bromberger's definition. For, firstly, there is a way of converting questions about the world into questions about a look up table. For example:

Q_o What is the world's highest mountain?

Q_m What words appear opposite "What is the world's highest mountain?" in the table?

[68] Note that the pragmatic conception of models still leaves open the possibility that the real Thunderbird 2 could, under certain circumstances, serve as a model of the small scale replica on my desk. For example, if I want to answer a question about the model Thunderbird 2, but do not have access to it, because I am marooned on Tracy Island, I might be able to use the real Thunderbird 2 to help me. The point is that, *in a given context*, the relationship between a model and the object it represents is, from the pragmatic point of view, asymmetrical, not that this relationship can never be reversed.

Secondly, as the above example illustrates, Q_m is generally a lot easier to answer than Q_o, so look up tables qualify as models according to the relevant pragmatic criteria. And finally, having obtained an answer, A_m, to Q_m, it is a simple matter to answer Q_o:

A_m "Mount Everest" appears opposite the words "What is the world's highest mountain".

A_o Mount Everest is the world's tallest mountain.

An obvious response to this difficulty is to complain that a look up table is, in some sense, "gappy". Although we can use one to answer *some* questions about a particular domain, we cannot use it to answer *all* of those questions. In contrast, a model of Thunderbird 2 apparently enables me to answer any number of questions about its object. The problem with this response is that models can be "gappy" too. For example, unless it is very sophisticated, my model of Thunderbird 2, won't help me to determine the cruising speed of the real thing. True, the gappinness that many models exhibit seems different from the gappinness of a typical look-up table, but, unless we can say what the difference is, this observation is of limited use.

In any case, even if we could provide a satisfactory account of the above matter, our objector could simply replace the gappy look-up table with an *oracle* that *does* have all the answers; or all those, at least, that relate to a given domain of inquiry[69]. Such an oracle would still satisfy Bromberger's definition, since, firstly, there is a systematic way of converting questions about the world into questions about the oracle:

Q_o What is the world's highest mountain?

Q_m What does the oracle say when asked, "What is the world's highest mountain?"?

Secondly, having an oracle answer one's question is obviously a lot easier than answering it oneself, so the pragmatic criteria for being a model are satisfied. And finally, having obtained an answer, A_m, to Q_m, it is a simple matter to answer Q_o:

[69] Once we have the idea of an oracle, we can treat look up tables as special cases: in effect, they are just imperfect oracles.

A_m The oracle's reply to the question, "What is the world's highest mountain?", is "Mount Everest"

A_o Mount Everest is the world's tallest mountain.

The problem is that oracles look even less like models, as usually conceived, than look up tables do!

So, Bromberger's definition of a model needs to be modified to exclude both look up tables and oracles. The best way to do this, I believe, is to invoke the distinction between being told the answer to a question, and discovering the answer for oneself. In the above example, the reason that I can answer the question, "What is the world's highest Mountain?" by referring to the look up table is that, at some point, someone has answered this question *without* referring to a table or any similar device. The utility of a look up table derives, ultimately, from the fact that all of the questions it lists have already been answered by someone else. The same can be said of oracles: before I get an answer to my question from the oracle, it must somehow answer the question itself. Again, someone or something *must* know the answer before I can.

This is not the case with genuine models. Consider, for example, the model of Thunderbird 2. I can use it to answer questions such as, "How many windows does Thunderbird 2 have?", and, "How many doors does Thunderbird 2 have?". These questions will probably have been answered directly, by counting the windows and doors of the real Thunderbird 2, before constructing the model. But now, compare the question, "Does Thunderbird 2 have more windows than doors?". It is highly unlikely that this question was *asked*, let alone answered during the construction of the model, and in any case, it is not necessary that it was. Nevertheless, given the model of Thunderbird 2, it is a simple enough question to answer.

Still, it is not quite correct to say that the difference between a model and an oracle, or look up table, is that it is possible to use the former, but not the latter, to answer questions about its object that were not necessarily answered during its construction. To see this, consider the following, very simple, look up table:

Question	**Answer**
How many doors does Thunderbird 2 have?	Two
How many windows does Thunderbird 2 have?	Twenty

Although the above table does not *directly* answer the question, "Does Thunderbird 2 have more windows than doors?", the information it *does* supply can be used to work out the correct answer. Furthermore, this question need not have been answered explicitly during the construction of the look up table. So, the above table can be used to answer a question that was not necessarily answered during its construction. Likewise, we could put the two questions in the above table to a suitable oracle, and use the answers it provides to answer the third question, without directly asking the oracle for the answer. These considerations show that the proposed way of distinguishing between models and oracles, or look up tables, is not quite correct.

To disarm this objection, we need to distinguish between using a look up table *qua* look up table, and using it as a source of premises that can be used in other inquiries we may wish to pursue. Use of a look up table *qua* look up table is a quite mechanical procedure that involves locating the question to be answered in the table, and then, if it is there, noting the answer that appears opposite it. If we restrict use of the above table in this way, then it *cannot* be used to answer the question, "Does Thunderbird 2 have more windows than doors?". Conversely, the only reason it *can* be used to answer this question, in the example, is that it is not used in this way, but is used, instead, as a source of premises; and these, in turn, are used to arrive at the correct response.

Similar remarks apply in the case of an oracle: its *proper* use – the type of use that distinguishes it *as* an oracle – involves putting questions to it directly, and accepting whatever answer is given. But, the oracle in our example is not used in this way; rather, it is used as a source of premises that are the starting point for a subsequent inquiry. Of course, the oracle *could* be used (*as* an oracle, that is) to answer the question, "Does Thunderbird 2 have more windows than doors?"; but then this question has been answered by someone, or something, else, and is not a counterexample to the way that I have proposed to distinguish oracles from models.

There is one further complication to bear in mind before we leave this topic. Sometimes, a model, like a look up table or oracle, may be used to supply the premises for a subsequent inquiry. In fact, this is just what happens in the above example: once we use the model to answer the questions, "How many windows does Thunderbird 2 have?", and, "How many doors does Thunderbird 2 have?", we can work out the answer to the question, "Does Thunderbird 2 have more windows than doors?", without referring to the model. This suggests that we need to stipulate the way in which a model may be used to answer questions, for the purposes of the distinction we are trying to draw – as is the case with look up tables and oracles, the model must be used in a way that is distinctive of it *qua* model. Intuitively, this involves performing experiments with it, taking measurements, and observing the model in any number of other ways.

To summarise our discussion of the nature of models, we began with Bromberger's definition:

> M is a model of O, relative to a set of triples [Q_m, Q_o, T*] *iff* in each triple, Q_m is a set of questions about M, Q_o is a set of questions about O, and T* is an algorithm that translates any answer to a member of Q_m into an answer to a member of Q_o, and *correct* answers to the former into *correct* answers to the latter.

Our examination of a number of objections to this definition suggests that the following may be nearer the mark:

M is a model of O, relative to a set of triples [Q_m, Q_o, T] *iff*:

1. In each triple, Q_m is a set of questions about M, and Q_o is a set of questions about O

2. T represents a set of procedures (algorithmic and/or heuristic), by means of which it is possible, though not necessarily easy, to translate any member of Q_o into a member of Q_m, and any answer to a member of Q_m into an answer to the corresponding member of Q_o.

3. T translates correct answers to members of Q_m into correct answers to members of Q_o.

4. M enables us to answer at least one member of Q_o that *could* have been left unanswered during M's construction.

Of these conditions, the fourth is critical, as it distinguishes a genuine model from an oracle or look up table, as long as it is understood that each of these devices is being used in the way that is distinctive of it.

What does all of this have to do with the *Meno* dialogue, or with the nature of understanding? To see the connection, recall that, in the previous section, we concluded that, very roughly, minimal comprehension is demonstrated when new information is taken entirely on trust, rather than being subjected to critical reflection, in the light of one's other beliefs. Certain episodes in Plato's dialogue led us to believe that this is precisely the attitude that Meno adopts towards the opinions of others, treating them as things to be memorised and reproduced, more or less *verbatim*, should a suitable occasion arise. But consistently following this strategy is bound to have a debilitating effect upon Meno's intellect, for it leaves him unable to answer any question that has not already been answered by somebody else. In the light of our recent discussion of the difference between models and look up tables, we can now say that, because he never gets beyond the level of minimal comprehension, Meno becomes reliant upon a kind of internal look up table, which acts as the repository of other people's opinions. For example, when asked, "What is virtue?", we can almost imagine Meno asking himself, "How did Gorgias reply to that question?", and answering accordingly[70].

So, consistently responding to new information with no more than minimal comprehension leads to reliance upon a kind of internal look up table[71]. This being an unsatisfactory state of affairs, it is natural to hope that consistently making the effort to comprehend new information to a *greater* than minimal degree leads to a more worthwhile outcome. Could it be, then, that the tenacious exercise of non-minimal comprehension enables one to construct a model of whatever it is that the information is about? As this section has begun to demonstrate, a model contains a large amount of information about the object it represents, in a coherent, unified form. This conception of a model is already reminiscent of our intuitive ideas about understanding, so the idea that non-minimal comprehension facilitates model construction is certainly plausible.

[70] Not that we need to imagine Meno going through this procedure explicitly; the important point is that Meno responds to questions *as if* he is using an internal look up table in this way; in other words, a look up table is an adequate model of Meno's cognitive processes!

[71] Or, possibly, several internal look up tables.

If the above hypothesis *is* correct, this implies something about the nature of *static* understanding, since:

1. The consistent exercise of non-minimal comprehension with respect to information about a given domain, enables one to construct an adequate model of that domain[72].

2. The consistent exercise of dynamic understanding in a given domain may reasonably be expected to lead to static understanding of that domain.

3. Dynamic understanding is exhibited whenever an episode of non-minimal comprehension occurs.

4. *Therefore*, one has static understanding of a given domain to the extent that one has an adequate model of it.

The remainder of the thesis is devoted to exploring the idea that static understanding of a given domain is a matter of having an adequate model of it.

3.6 Summary

The idea that understanding is the goal of inquiry sounds appealing, but is of limited use, unless we have a clear account of what understanding *is*. In this chapter, we have begun to develop such an account.

Our starting point was to distinguish between dynamic and static understanding: the former is episodic, occurring at a specific time, whilst the latter is a state of the individual that persists over a period of time. The two aspects of understanding are related, in that we may reasonably hope that a series of episodes of dynamic understanding will lead to static understanding, in the long run. The core of the chapter was then devoted to developing an account of dynamic understanding, concluding that it is exhibited whenever an individual responds to new information with non-minimal comprehension, as defined at the end of section 3.4.

Finally, in section 3.5, we introduced the concept of a model, and distinguished models from look up tables and oracles. Towards the end of the section, we decided that a series of

[72] Or, at least, it is reasonable to *hope* that it will.

episodes of non-minimal comprehension only ever leads to the construction of a kind of internal look up table, rather than a model of the domain in question. This suggested that a series of episodes of *non*-minimal comprehension (and, therefore, dynamic understanding) within a given domain, may enable one to construct a model of that domain. Since a series of episodes of dynamic understanding also leads to static understanding of the domain concerned, this implies that one has static understanding of a given domain, to the extent that one has an adequate model of it.

The remainder of the thesis is devoted to exploring this idea in more detail: In chapter 4, I argue that there is an analogy between spatial orientation and understanding, and that this provides good evidence for the claim that we understand a given domain to the extent that we have an adequate model of it. And in chapter 5, I show how this claim makes sense of current debates about the nature of scientific theories and scientific explanation, and argue that this constitutes further evidence in its favour.

Chapter 4: Understanding and Spatial Orientation

4.1 Introduction

Towards the end of chapter 3, I suggested that *static* understanding[1] of a given domain is a matter of having a model of it. In this chapter, I argue that there is an analogy between spatial orientation and understanding, and that this provides further evidence for this claim[2]. That this is a reasonable avenue to explore is suggested by the prevalence of spatial metaphors in the language we use to describe our inquiries:

> Intellectual disciplines are divided into different *areas* or *fields* of study. Researchers occupy and develop *positions* within the field, and, if they are lucky, find themselves *making progress*, or *getting somewhere*. Perhaps an inquiry will reach a satisfactory *terminus*, and it will be possible to describe a *line of thought* which *leads step by step* to the conclusion *arrived at*. Sometimes there are *obstacles along the way* which the researcher needs to *overcome*, or *find some way around*. If things go really badly, the result is often a feeling of being *hopelessly lost*, or of *rambling aimlessly*, without a *sense of direction*...[3]

Section 4.2 makes a start by introducing the concept of a spatial problem. After considering what options are available to an individual facing such a problem, it concludes that some form of spatial orientation is most satisfactory. In arriving at this conclusion, the section notes, in passing, some interesting parallels with various responses to the demands of rational inquiry. Section 4.3 examines the nature of spatial orientation in more detail, and reveals it to be a multi-faceted phenomenon. Section 4.4 looks at some situations where we are inclined to say that understanding is present and shows that they exhibit the same features that are characteristic of spatial orientation. Finally, section 4.5 uses these similarities as the basis of an argument from analogy, the conclusion of which is that understanding a given domain is a matter of having a model of it.

[1] As opposed to *dynamic* understanding; see the beginning of chapter 3, p80ff for an account of this distinction. As the remainder of the thesis is concerned primarily with static understanding, I shall omit the qualification from now on.

[2] This line of inquiry was suggested to me by Neil Cooper's remark in "Understanding", p4, that understanding is the "geography of knowledge". He also makes use of an analogy between maps and understanding in "The Epistemology of Understanding", p10.

[3] Metaphors with spatial connotations are not the only ones found in the language we use to describe our intellectual endeavours. We talk, for instance, about *digesting* new information, and about *regurgitating* what we've learned. However, I believe that the language of spatial orientation is the source of the most extensive, systematic network of metaphors in this area.

4.2 Spatial Problems

What's the shortest route from my house to the philosophy department? Where's the nearest Italian restaurant? How do I get to the railway station from here? These are examples of what I shall call *spatial problems*[4]. The possibility of such problems arises, at least partly, for two reasons. Firstly, everyone, and everything, with the possible exception of abstract objects, has got to be *somewhere*, usually, as far as we're concerned, on or near the surface of the Earth. Secondly, it is impossible for two people or things to occupy exactly the same location at the same time, so *separation* or *distance* between people, places or things is inevitable. Although some of the things we want are ready to hand, many more are out of reach, *elsewhere*, and we must find some way of getting them *here*.

There are at least three types of response I can make to a spatial problem[5]:

1. Reject the problem
2. Replace the problem (usually with one that is more tractable)
3. Attempt to solve the problem

Let me illustrate these alternatives with the aid of an example. Suppose that I decide to treat myself to a jaffa cake as a reward for working hard on my Ph.D, but am not sure where I can lay my hands on one. The following responses to this problem are available to me, the first and second corresponding with alternatives 1 and 2 in the above list, and the remaining four being strategies, of varying degrees of sophistication, for actually solving the problem:

Reject the Problem: One possibility is just to give up on the idea of getting a jaffa cake. This has the virtue of simplicity and ease of implementation, but it is hardly satisfactory, as I do not get what I want[6]. Unless I can find a way of moderating my desire, I shall spend the rest of the afternoon nursing an unsatisfied longing for a jaffa cake, and this could have a disastrous effect on my Ph.D.

[4]The following discussion is taken from *Maps in Mind: Reflections on Cognitive Mapping*, by R. M. Downs and D. Stea.

[5] Or any kind of problem, for that matter.

[6] Though this strategy may not be so bad if I can to modify my preferences so that I no longer desire the jaffa cake.

Likewise, giving up is a possible response, however inadequate, to the demands of any inquiry[7]. Adopting it may be accompanied by despair, if the craving for knowledge persists, or relief, if the hard work associated with genuine inquiry is felt to be too much of a burden. It is interesting to recall, in this connection, that Socrates objects to Meno's paradoxical argument on the grounds that accepting its conclusion would promote epistemic apathy:

> We ought not then to be led astray by the contentious argument you quoted. It would make us lazy, and is music in the ears of weaklings. The other doctrine [i.e. the Theory of Recollection] produces energetic seekers after truth[8]...

Replace the Problem: A second option is to make the original problem more tractable by *substituting* some other form of satisfaction for the absent object of desire. For example, I decide to eat the apple I keep in my desk drawer, rather than go looking for a jaffa cake. This is an improvement on the first response, since, in some sense, my desire is satisfied[9], but it is still not satisfactory, as it will not *generally* succeed. What happens, for instance, when the apple isn't there either, or when only a jaffa cake will do?

In general, this response involves replacing a hard question, Q_0, with a second question, Q_1, that is thought to be easier to answer[10]. As our discussion of models in the previous chapter brought out, this is not always a bad strategy: sometimes solving one problem opens up the possibility of an indirect solution to the original difficulty. However, insofar as this response is adopted not merely as a means to an end, but as an end in itself, it is an inadequate solution to the demands of inquiry.

As explained above, the remaining four possibilities all at least *try* to solve the problem of the missing jaffa cake.

Delegation: The first strategy for solving the jaffa cake problem is one that most of us are familiar with from childhood: "Mum, will you fetch me a jaffa cake, please". Instead of trying to solve the spatial problem myself, I delegate the task of solving it to an

[7] Not that rejecting the problem is always an inadequate response; it seems, for instance, perfectly rational not to look for a way of turning jaffa cakes to diamonds.

[8] *Meno*, 81d.

[9] Not my original desire for a jaffa cake, admittedly, but perhaps my desire for a reward for all my hard work.

[10] Another possibility is to lower the standards by which candidate answers to Q_1 are judged. For example, instead of insisting that the answer be true, or rationally justified, we might decide to settle for an answer that makes us feel good. This response may not be coherent though: arguably the

intermediary. This strategy has the advantage of addressing the problem as it is given, and of doing so with a fair degree of generality[11]. However, as we grow older it becomes a less and less reliable way of getting what we want; other people have lives too.

In chapter 3, we saw that undue reliance on other people can be debilitating in the general case also. This is precisely the strategy that Meno adopts in all of his inquiries: get somebody else to do all of the work, and then accept their answer without question[12]. This accounts for his unease when Socrates refuses to give a direct answer to his question about the origin of virtue. However, the strategy of delegation has a respectable as well as a disreputable aspect. Whenever we consult a work of reference in order to answer a question, we rely upon the work of others in a way that seems entirely innocuous. So the point is not that it is *never* acceptable to benefit from the intellectual labour of other people, just that it is *not always* acceptable. A community in which everyone always chose delegation in response to the demands of inquiry, would not be able to successfully conclude *any* of its inquiries, due to a regress of delegation.

As previously noted, delegation differs from the first two responses in addressing the jaffa cake problem as it stands. However, this strategy differs from remaining ones described below in that I do not solve the problem *myself*. Instead, I rely, to a greater or lesser degree (a continuum of cases exists), on the efforts of someone else. From now on, we're on our own.

Locational Luck: The second strategy for solving the jaffa cake problem relies not upon other people, but upon what might be called a *windfall*. By chance, I might be right next to a packet of jaffa cakes when the craving strikes, in which case, all I have to do is engage in some elementary, and pleasurable, hand-eye-mouth co-ordination. However, unless I am really lucky, it is unlikely that everything I want will be in my field of vision[13], so, in practice, this strategy will tend to degenerate into either rejection or replacement of the problem.

concept of inquiry necessarily involves the idea that its aim is to represent how things stand accurately.

[11] But note: it's not clear that this strategy works where the problem is a matter of *being* in a certain place, at a certain time; someone else could hardly do that for me.

[12] See chapter 3, pp106-108 and p116.

[13] Or, more precisely, within one of any number of fields of vision I can access by swivelling my head.

In the case of inquiry in general, this strategy is akin to adopting the attitude of the ancient sceptic, and allowing ourselves to be guided entirely by the appearances in forming our beliefs, without delving further into their sources and significance. An individual who follows this strategy is willing to endorse only those opinions which lie, as it were, ready to hand[14].

Relying on locational luck to solve one's spatial problems may sound like a trivial solution, but it highlights something very important about the nature of such problems. For success using this strategy depends upon the jaffa cake being within effective range of at least one of my sensory modalities. And the fact that we find it natural to refer to success under those circumstances as a windfall indicates that it is not the norm. The reason for this is that our senses have certain limits beyond which they are unable to supply us with information. However, our environment does not consist of just that region defined by these limits[15], but extends indefinitely beyond them. Hence, at any given time, I am aware of only a relatively small fraction of all the information about my environment that is, in principle accessible to me[16]. This is the underlying reason for the inadequacy of our second strategy for solving the jaffa cake problem. In order to remedy this defect, we must introduce an element of *exploratory locomotion* into our response to the spatial problem. The final two strategies do just that[17].

Random Walk: The third strategy for locating a jaffa cake is to go on a *random walk* through the house until I get within sensory range of a jaffa cake. The degree of randomness can vary, of course. For example, I might start looking in an entirely random fashion, but with the intention of trying not to check the same place twice. Introducing this element of exploratory locomotion diminishes the element of luck that infected the previous strategy up

[14] This is a very crude depiction of the positive aspect of the Pyrrhonist way of life. The interested reader should consult Sextus Empiricus, *Outlines of Scepticism*, B. Mates, *The Sceptic Way* (a new translation of Sextus by Mates with an excellent introduction and commentary), and C. Hookway, *Scepticism*, chapters. 1-2.

[15] We usually do not think of the effective range of our senses as delimiting a region. This, I speculate, is because most of us are mobile most of the time, so the delimited region is changing continuously.

[16] Furthermore, even my access to *this* bit of the environment is limited: I might not be able to see the jaffa cakes *from here* if they are hidden behind the breadbin, for example. Unless I live in a chic minimalist apartment, even my immediate environment is *cluttered*, so that certain information is denied me as one thing obtrudes in front of another.

[17] Note that the last two strategies represent two ends of a continuum. In the first, I have *no* information about the location of the jaffa cakes, whereas in the second, I know where the jaffa cakes are, and simply have to make may way to them. In practice, there are any number of strategies that lie between these two extremes. For example, I might know that the jaffa cakes are either in the lounge or in the kitchen, which narrows down my search a lot, but then have to rely on random walking when searching *within* these two rooms.

to a point, but obviously does not eliminate it altogether. For this reason, random walking must be judged inadequate as a general strategy for solving spatial problems. Likewise, conducting our inquiries in a random fashion simply leaves too much to chance to be a good strategy in the general case.

Orientation: The fourth, and best, strategy is to use my *knowledge* of where the jaffa cakes are[18]. In this instance, there are two distinct stages involved in the execution of this policy. First, I *locate* the jaffa cakes; then I head for that location, relying on my knowledge of the spatial layout of the house. As I suggest later, one way of putting this is that I rely upon a kind of map, a *cognitive* one in this instance, of my environment to help me solve this particular spatial problem. The key feature of this strategy is that it supplements perception with *cognition*; supplements the sensory information I receive now, with all of the relevant information I have at my disposal from memories of other trips around the house. Accessing this information enables me to form a *plan* that will get me to the jaffa cakes quickly and efficiently instead of having to rely upon finding them by chance.

In general, the counterpart of this strategy is the kind of intelligent inquiry that Socrates and Peirce believe is both desirable and possible.

Section 4.3 explores the nature of spatial orientation in much more detail.

4.3 Aspects of Spatial Orientation

Spatial orientation is a relationship between an organism and its environment[19]. Loosely speaking, an organism is orientated with respect to an environment when it "knows its way around". Obviously, spatial orientation is a matter of degree, ranging from complete ignorance of the environment to thorough familiarity with it. What is not so obvious is the number of dimensions along which the degree of orientation can vary. In this section, I describe some of the main ones. I begin by introducing a relatively formal definition of what I call *maximal* orientation that highlights three of the main factors involved, before describing other aspects of spatial orientation that this definition implies.

[18] Or where they would be if there were any. I might, for example, know where the biscuit tin is, without knowing if there are any jaffa cakes in it. In this case, making my way to the tin is my best *hope* of finding a jaffa cake in the house, in the sense that if any strategy can succeed, kit can.

[19] For the sake of simplicity, think of an environment as a set of places, and a network of linking paths. My environment, for example, is (part of) the city of Sheffield: it includes such locations as my house, the Arts Tower, and the railway station, and the network of streets that connects them all.

To facilitate the discussion, let's introduce a simplified model of the phenomenon we're interested in. Consider the situation of an organism, O, which has to find its way around a region, R. Assume that O has but a single mode of orientation; that is, there is just one sensory channel - sight, let us say - via which environmental information reaches the organism. The region R contains a set of landmarks, L_1, L_2...,L_n linked by paths. Assume that O can recognise all of these landmarks by sight, and is able to distinguish any one of them from any of the others[20]. Let us define a spatial problem as the task of reaching a target landmark in the domain, L_t, from a specified starting point L_s. If O knows absolutely nothing about its environment, we say that it is *completely disorientated.* In such circumstances, the organism's only recourse is to random walking, and any success will be purely accidental. This state lies at one end of the orientation continuum, and may be defined thus:

An organism, O, is completely disorientated in an environment, R, iff, there is no spatial problem within R which it can solve with a success rate greater than that which could be achieved by random walking.

Conversely, we can attribute *some* degree of orientation to the organism as long as there is at least one spatial problem within the domain which the organism solves with a success rate greater than that which could have been achieved by random walking.

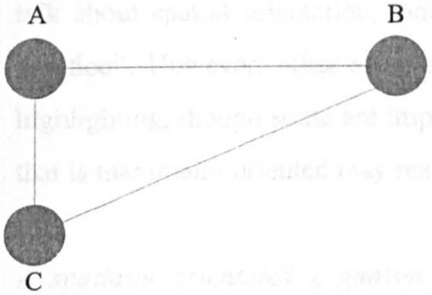
Spelling things out in this way brings out the fact that degree of orientation can vary along at least two different axes. Firstly, it is a matter of degree how *often*, with what success rate, the organism can successfully complete a given route. Secondly it is a matter of degree how many *different* routes the organism can complete with a degree of success greater than the random walk rate. Put slightly differently, orientation is manifested both in the ability to solve many spatial problems, and in the fact that success is not accidental, but rather is attributable to the organism's intelligent exploitation of information about its environment. At this point, we can make a first attempt at formulating a definition of*maximal orientation*, the state that lies at the opposite end of the continuum to complete disorientation. Such a definition would look something like this:

An organism, O, is maximally orientated in an environment, R, iff it can successfully solve all spatial problems within the domain every time.

[20] Hence, there are no situations in which O can only tell it is at a given landmark by the route it has taken to get there.

However, as the following example suggests, we can do better than this:

Figure 4.1: Efficiency and Maximal Orientation



The above diagram depicts an extremely simple environment consisting of just three locations, A, B and C, each of which is linked to the other two by a straight path. Suppose that the organism, O, can successfully get from A to B, B to A, B to C, and C to B every time, using the marked paths. It can also travel from A to C and back again every time. O satisfies the above definition of maximal orientation, with respect to this simple environment, since it can solve all possible spatial problems that arise within it, all of the time. But suppose that in travelling from A to C, O goes via B, and likewise for the reverse trip! This is an extremely inefficient way of proceeding, given that a much shorter, direct route is available. Whether O is aware of the direct route AC or not, such behaviour is less than satisfactory from the point of view of spatial orientation.

This example shows that there is another dimension along which the degree of orientation can vary, that of *efficiency*. Hence, we ought to modify our definition of maximal orientation along the following lines:

An organism, O, is maximally orientated in its environment, R, iff it can find the most efficient solutions to all of the spatial problems within the domain every time.

A brief word is in order about the meaning of "efficiency" in the above definition. The organism seeks to maximise benefits and minimise costs in making a journey. Benefits such as scenic beauty, or costs such as danger of being mugged, or how hilly the terrain is, may influence the choice of a route in real life, but I suggest that we ignore these in our simplified model, and assume that travelling time is the only relevant factor. Let us also assume that travelling time is directly proportional to the distance involved. In these circumstances, the most efficient solution to a spatial problem is the one that takes the shortest route from starting point to destination.

The above definition of maximal orientation captures the essence of what we mean when we talk about spatial orientation, though it describes an ideal state that is seldom realised in practice[21]. However, other aspects of our intuitive notion of spatial orientation are worth highlighting, though some are implicit in the above definition, in the sense that an organism that is maximally oriented may reasonably be expected to exhibit them:

A spatially orientated organism responds flexibly to spatial problems. There may be occasions when a preferred route - the most efficient one if the organism is maximally orientated - is not available, because, for example, the path is blocked. In these circumstances, the only option, apart from giving up, is to take an alternative route. The extent to which an organism is able to take advantage of such alternative routes as the need arises, indicates the degree to which it is orientated with respect to its environment. Flexibility of response depends upon the organism being aware of two or more routes between the same pair of locations, and, in addition, being able to *recognise* that any one may be substituted for any of the others, if necessary.

A spatially orientated organism responds robustly to spatial problems. When an organism follows a route it relies on a variety of sensory cues to monitor its progress, and to inform future decisions about which path to take. The absence of any one of these cues, a distinctively shaped rock used as a landmark, for example, is a potential source of disorientation. To the extent that an organism is able to cope with these disruptions, it is better orientated in its environment. Put another way, a tendency to become disorientated when a familiar landmark is moved or hidden is a sign that the organism, being overly reliant upon a single piece of environmental information, is not adequately orientated in the first place.

A spatially orientated organism can reverse its solutions to spatial problems. If O is able to find its way from A to B, then, it ought, if it is well orientated, to be able to retrace its steps and follow the route from B to A. This is just a special case of the requirement, contained in

[21] Unless, perhaps, the environment with respect to which orientation is being measured is relatively small.

the definition of maximal orientation, that O has the ability to solve many spatial problems, but it is worth highlighting because, as we shall see in the next section, there are interesting parallels in other areas.

Like the aspects of spatial orientation highlighted in the definition of maximal orientation, all of the above additional aspects are matters of degree: strictly speaking, we should say that the more flexibly (or robustly, and so forth) an organism can respond to spatial problems, the better orientated it is with respect to that particular environment.

To summarise the discussion of this section, spatial orientation turns out to be a complex phenomenon, with at least the following (not necessarily distinct) aspects:

- Ability to solve *many* spatial problems.
- Success in solving spatial problems is *non-accidental*.
- Solutions to spatial problems are *efficient*.
- Solutions to spatial problems are *flexible*.
- Solutions to spatial problems are *robust*.
- Solutions to spatial problems are *reversible*.

4.4 Understanding as Orientation

In this section, I show that the features of spatial orientation listed at the end of the previous section may be found in other domains, in circumstances where there is what we would intuitively describe as understanding.

4.4.1 Ability to Solve Many Problems

Just as the ability to solve many spatial problems is indicative of a high degree of spatial orientation, so the ability to solve many problems within a given domain is a sign of understanding. For example, we normally attribute to a professor of sub-atomic physics a greater degree of understanding of the discipline than an undergraduate. And this judgement is based partly on the fact that the professor can solve more problems in that domain than the student.

However, this is a little crude as it stands: maybe the professor can solve more problems than the student because she knows more facts about sub-atomic particles, rather than because her understanding of the subject is any greater. So, we need to stipulate that the problems do not require for their solution facts known only to the professor; they should be ones that any intelligent undergraduate *could* solve[22], given the basic principles of sub-atomic physics. Even under these conditions, we would confidently expect the professor to outperform the student[23]. This more sophisticated way of describing the example makes it clear that understanding is not a matter of acquiring more and more information, but of using the information one already has well.

4.4.2 Success in Problem Solving is Non-Accidental

Our concept of spatial orientation also demands that the success of the organism in finding its way around be non-accidental. At the very least this means that its success rate is greater than that it could achieve by random walking. In other domains too, a problem solving rate greater than that predicted by chance is regarded as a mark of understanding. For example, the probability of correctly guessing the answer to a single "true or false?" question taken from a first year logic paper is 1/2. Therefore, a score of 50% in such a test does not provide good evidence that the student has really understood the lecture material. However, correctly answering all twenty "true or false?" questions on the paper is good evidence of understanding, because the odds against achieving this level of success by guesswork are so great[24].

A slight complication arises if we suppose that the student that has understood the lecture material nevertheless makes a single mistake in working out each of her answers, with the result that she scores zero overall. Obviously, this is far worse than she could have done by just guessing, but we are unwilling to say that she lacks understanding. We can deal with this difficulty by invoking the well known distinction between competence and performance[25]. *Ex hypothesi*, the student has a good grasp of first year logic, and is therefore competent to answer the questions in the test. Under normal circumstances she would score highly, and this successful performance would provide evidence for her competence. So, in the unfortunate case where the student makes a mistake in answering every question in the test, this is an unlucky failure to perform, rather than a reflection of her underlying competence

[22] One way of putting this is that the student could be led to the solution of the problem via a series of well chosen questions to which they already know the answer. The similarity with Socrates' method is obvious.

[23] Still, it might be said that the professor has more *practical* knowledge than the student. I am not inclined to object to this claim: understanding may indeed be a form of *practical*, as opposed to theoretical, knowledge. For more on the distinction between practical and theoretical knowledge, see A. W. Moore, "Ineffability and Reflection: An Outline of the Concept of Knowledge".

[24] Just over a million to one, in fact.

[25] The distinction is made by Chomsky in "On the Nature, Use and Acquisition of Language".

(though if the student consistently failed to perform, we might have to revise our estimate of her competence).

One way of picking up this kind of case would be to require students to justify their answer to any particular question, in order to test their understanding; for it is very unlikely that a student will be able to come up with a reasonable sounding justification by chance. In either case, we look for reassurance that the student's success is non-accidental[26], in order to satisfy ourselves that understanding is present.

4.4.3 Solutions to Problems are Efficient

The ability to find efficient solutions to problems within a given domain is a hallmark of understanding, just as spatial orientation is partly manifested in the ability to find efficient routes. In the latter case, we assumed, reasonably, I think, that the efficiency of a route was inversely proportional to its length. However, in other domains, there is no ready made efficiency metric. In particular we should beware of confusing *speed* with efficiency. A high-powered Pentium PC will deliver the answer to any given problem much more quickly than a ZX Spectrum, even if both computers use exactly the same algorithm. But to conclude that the Pentium PC is more efficient on this basis is akin to claiming that a tortoise can never be as well orientated, within a given environment, as a hare, simply because it is physically incapable of moving as fast.

This suggests that, in general, the efficiency of a solution is related to the algorithm it employs, rather than to the speed with which that algorithm, once chosen, is executed. The problem is that this sort of efficiency is difficult to describe precisely. One suggestion is that the efficiency of an algorithm is inversely proportional to the average number of steps it will execute before finding a solution. However, in order to use this idea, it is necessary to specify in advance what operations are permitted; and any such choice will be arbitrary, to a certain extent.

However, there is no need to get bogged down in these issues; for it is obvious that we *do* have standards for judging when one solution to a problem is more efficient than another,

[26] There is another subtlety to be aware of: it seems that the only way of testing whether success is non-accidental is to get the student to answer *more* questions; this blurs the distinction between the first two aspects of understanding (ability to solve many problems; problem solving success is non-accidental).

even though these might be hard to formulate clearly. Two examples will illustrate the point:

Example 1: Consider Fermat's last theorem: $x^n + y^n \neq z^n$, for all $x, y, z \in N$, and all $n > 2$. Fermat claimed to have a proof of this theorem which he could not fit in the margin of the book he was annotating at the time. Notoriously, any such proof eluded generations of mathematicians, until Andrew Wiles of Princeton discovered one recently. Although it is a remarkable achievement, Wiles' proof runs to hundreds of pages, and makes use of mathematical concepts that would not have been available to Fermat. If Fermat really did have a proof of the theorem, it's a fair bet that it was much shorter and simpler than the one we have. If that is the case, then Fermat found a more efficient solution to this particular problem than we have managed to do so far.

Example 2: Genghis and Attila are challenged to solve the equation: $(x-5)(x+2) = 0$. Genghis knows how to multiply out bracketed algebraic expressions, and does so, obtaining the equation: $x^2 - 3x - 10 = 0$. He also knows the formula for the general solution of a quadratic equation, namely:

If $ax^2 + bx + c = 0$, then $x = (-b \pm \text{sqr}\ (b^2 - 4ac)) / 2a$

Genghis plugs the values, $a = 1$, $b = -3$, and $c = -10$, into the above expression and thereby discovers that x is equal to either -2 or 5.

Most of us are inclined to say that Genghis has gone the long way round, just because he is used to the methods it involves. He is like the individual in the previous section who travels from A to C via B, rather than taking the direct route. For, in this example, there is a much more efficient way of proceeding. Attila examines the original equation, and, although he is fully aware of the formula that Genghis uses[27], also recognises that if the product of two numbers is zero, then at least one of the numbers is also zero. He then reasons as follows:

If $(x - 5)(x + 2) = 0$, then either $x - 5 = 0$, or $x + 2 = 0$

Therefore, either $x = 5$, or $x = -2$

We may assume that this reasoning does not use any information that is unavailable to Genghis, in the sense that he would recall the implications of the product of two numbers being zero, if prompted. However, Attila connects this piece of information to the problem at hand, and Genghis doesn't, and in doing so opens up an efficient route to the solution. In short, Attila exhibits a greater degree of understanding than Genghis.

4.4.4 Solutions to Problems are Flexible

In the previous section we agreed that flexibility in one's approach to spatial problem solving is a matter not just of knowing two or more routes between the same pair of locations, but of being able to access that knowledge when it is needed; when, for instance, the direct route from A to B is blocked[28]. Such flexibility is often associated with understanding in other domains. In the example just discussed, for instance, Attila could demonstrate understanding of the problem by using Genghis' long-winded method if we perversely forbade him from taking the direct approach. The point is even clearer if we suppose that Genghis is unable to perform the complementary feat; he can't solve the problem unless he uses his original, inefficient method. Attila's understanding of the problem is superior to Genghis' because he has access to more than one way of solving it.

4.4.5 Solutions to Problems are Robust

An organism's response to a spatial problem is robust if it is not overly reliant upon any single piece of information about its environment. Something like the ability to cope with unexpectedly absent information is a sign of understanding in other domains. Consider the following problem, for example:

> A piece of rope is just long enough to circle the equator of the Earth. If the length of this rope is increased by just one metre, and it is again placed around the equator so that it is the same distance from the ground at all points, how far from the ground will the rope be. Assume the Earth is a perfect sphere.

Most people's instinct, upon encountering this problem, is to try to calculate the length of the extended rope. Since the rope forms a circle, it ought then to be possible to derive the radius

[27] Where Attila is not aware of this formula, the example is not so clear cut, since the idea is that Attila's reasoning reflects a deliberate choice on his part to find the most efficient solution to the algebraic problem.

[28] This turn of phrase is reminiscent of my description of models in the previous chapter. In the next section of this chapter, I argue that this is no accident.

of that circle using the formula, circumference = $2\pi r$, and from there subtract the radius of the Earth to obtain the solution. The problem is that the radius of the Earth is not given. So, unless we cheat by looking up this apparently vital piece of information in an encyclopaedia, the problem seems insoluble. But, as a matter of fact, we do not need to know the radius of the Earth to solve this problem. Instead, just call it R, and let the distance of the extended rope from the ground be d. Then, according to the problem:

Length of original rope = $2\pi R$

Length of extended rope = $2\pi R + 1 = 2\pi(R + d)$

So, $2\pi R + 1 = 2\pi R + 2\pi d$, in which case, $d = 1/2\pi$, or, a surprisingly large 16cm. An individual who solves this problem, despite the disconcerting absence of an apparently vital piece of information, exhibits a level of understanding that some people lack, even if they are familiar with all the relevant mathematics.

4.4.6 Solutions to Problems are Reversible

Spatial orientation may be exhibited by the ability of an organism to follow a route in the opposite direction to the one it is used to: if it can get from A to B, then it may reasonably be expected to be able, in most cases[29], to get from B to A. A similar ability is often a sign of understanding in a given domain. Take, for example, the problem of computing the volume, V, of a cuboid given the lengths of its sides, a, b, and c. This is easily done, as long as we are familiar with the simple formula $V = abc$. At first sight, the reverse problem is to determine the lengths of a cuboid's sides, given its volume. But this problem is insoluble, since there are infinitely many real number triples that give the desired result when multiplied together.

A better candidate for the reverse problem is this: given the volume, and the lengths of two sides of a cuboid, determine the length of the remaining side. In solving this problem, we *invert* many of the operations we used to solve the original problem; specifically, we substitute division for multiplication. This is not so different from the way a reverse route is traced by substituting left for right, and so forth. Furthermore, the ability to invert mathematical operations in this way is normally a sign that the student has understood the

[29] Some routes may be easy to traverse in one direction, but relatively difficult to reverse. For example, in a branching network of paths with a single node at the 'root', it will always be easy to get to the root, no matter where one is in the network, since all paths ultimately lead there; but travelling up from the root to a target branch is more difficult, on account of the choices one is forced to make at each junction.

subject matter. It is possible to imagine a situation where a student is only taught how to calculate volumes given lengths, using V = abc, and is then required to solve the reverse problem. If the student has really understood the lesson, and is familiar with some basic arithmetical principles, this is a simple matter[30].

To recap: so far, we have established that spatial orientation is a complex phenomenon with a number of different aspects. In this section, we have seen how similar features are often indicative of understanding in other domains. The next section is an attempt to account for both of these facts.

4.5 Maps, Routes and Models

So far in this chapter, we have established that spatial orientation and understanding have many features in common. In this section, I use these similarities as the basis for the following argument from analogy:

P1. Spatial orientation and understanding have many features in common.

P2. In the case of spatial orientation, we can account for the presence of these features by supposing that the organism has a map (usually a mental one) of its environment.

- *Therefore*, it is likely that the presence of the same features in cases of understanding may be accounted for by the existence of a map like representation of the domain in question.

In addition, I argue that a map is a model of the terrain it represents; this being so, the conclusion that understanding a given domain is a matter of having a model of it begins to look very plausible.

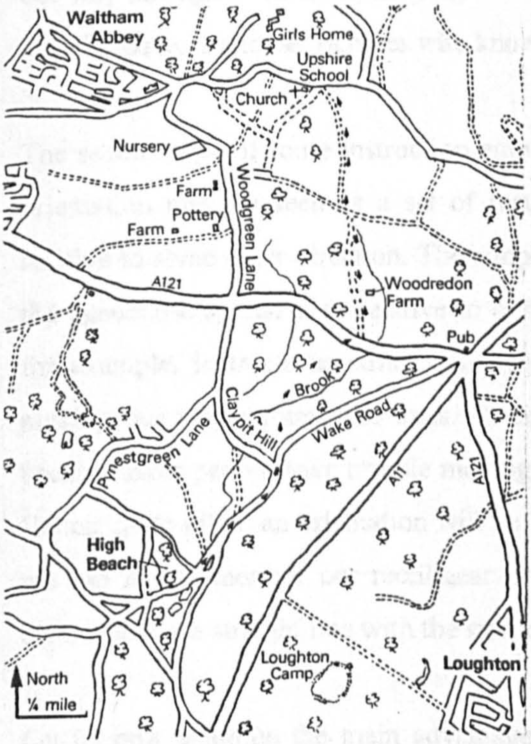
[30] The point is that the formula V = abc is *neutral* with respect to a range of possible applications. The importance of this point will become apparent in the next section on maps.

4.5.1 Maps and Spatial Orientation

In this section I aim to establish the second premise of the argument from analogy described above. In order to do so, we need to think about the properties of maps[31], and about how maps facilitate spatial orientation. I suggest that we address these issues by exploring the contrast between maps and another way of finding one's way around, *route descriptions*[32].

Consider the following example, taken from a book on country walks in and around London. A lost rambler approaches us with the question: "How do I get to Upshire?". Occasionally, through what we earlier called a piece of locational luck, Upshire will be visible from where we are now. It will then be possible to point out the village to the rambler, and leave her to make her own way there. More often than not, however, we will have to offer the rambler more elaborate assistance in the form of either a map, or a route description, as in figure 4.2.

Figure 4.2: Maps and Route Descriptions



Starting at High Beach, keep forward for 500 yards beyond the little tea hut to a point where the road divides after bearing leftwards. Go along the right fork for 30 yards, then strike left into the forest, passing between two beech trees. Continue ahead, leaving the road gradually on the right. Bear slightly right where the ground falls sharply away, and pick your own way through the trees, keeping the lower ground on your left. Keep on for almost 1/2 mile, when you may hear the sound of main road traffic ahead. After crossing one or two brooks, you will reach the road at the top of a hill near a bus stop. Almost opposite, you will find a metalled drive that leads through the trees to Woodredon farm. After passing the mansion on your left, the drive winds leftwards, then twists right again to emerge after 1/2 mile, at Upshire (the spire of St. Thomas Church is a useful landmark).

Let us begin by examining the route description in this example. Essentially, it consists of a list of stimulus-response-stimulus instructions which direct the rambler from one location along the route to another. For example, at the beginning of the walk, we are directed to keep

[31] Throughout this section, I shall concentrate on paper maps, as opposed to cognitive maps.
[32] This way of thinking about maps was suggested to me by the work, in *The Hippocampus as a Cognitive Map*, of the psychologists O' Keefe and Nadel. The example that follows, and much of the subsequent discussion, borrows heavily from that work; see especially pp80-89.

forward for 500 yards (response) beyond the little tea hut (stimulus) to a point where the road divides (stimulus). The stimuli referred to are primarily visual cues, but occasionally there is mention of sounds such as that of the traffic on the main road. There is no reason why stimuli appropriate to the modalities of touch and smell could not be used as well. If we were describing the route to a blind friend, for example, these might well have a significant role to play.

If we examine the route description carefully, we notice that some instructions place the emphasis on the stimulus and others emphasise the response: call the former*guidances*, and the latter *orientations*. A guidance directs the attention of the rambler to a particular landmark, such as the spire of St Thomas church near the end of the journey. The rambler is required either to approach the guide or to maintain some egocentric relationship to it, and is left to take whatever responses are appropriate. Guidances are usually relatively localised and stationary, but they need not be. For example, the route instructs us at one point to pick our way through some trees, keeping the lower ground on our left. And the best guidance of all, of course, is another rambler who knows the way; all we have to do is follow her.

The second type of route instruction emphasises the response rather than the stimulus. An orientation may be seen as a set of instructions for aligning the egocentric spatial axes relative to some other direction. The simplest case is one in which we are required to rotate the egocentric spatial axes relative to our present orientation. The route description given, for example, instructs us strike obliquely left into the forest, and to bear slightly right, picking our way through the trees. Orientations are usually somewhat difficult to follow because most people have trouble moving in a straight line without referring to visual cues. Hence, quite often, an orientation will be supplemented by a reference to a landmark which we can use to monitor our rectilinear progress. The best such landmark is one which is distant and in a straight line with the starting point and the goal.

Let us now examine the main advantages and disadvantages of route descriptions. Their most obvious feature, which must be regarded as a limitation, is that, as a general rule, they are written with a specific *goal* in mind. Hence, they can only be used to solve a rather limited range of spatial problems. Indeed, it might appear that a route description only enables us to solve a single spatial problem: that of getting from its starting point to its terminus. To see that this is not quite correct, consider a route description that tells us how to get from A to B via C. We could use this to solve the three spatial problems corresponding to the routes from A to C, A to B, and B to C. In our example, the goal is to get to the village of Upshire, but, as the route description links the present location with this goal via a series of intermediate places, it enables us to solve the spatial problem of getting from any one of these locations in the chain to any other[33].

Another disadvantage of routes is that they are *inflexible*. The instructions they contain must be followed in pretty much the order in which they are given, and this allows very little freedom of choice to the individual. However, route descriptions may incorporate a degree of flexibility occasionally. A trivial case occurs if all available paths are known to converge upon the destination, in which case, the route description can afford to let us decide which way to go. More common is the case where our immediate goal is to reach an intended guide, which all paths cross. For example, the route to Upshire instructs us to pick our way through the trees however we see fit since we are bound to reach the road eventually. Finally, a route may allow for various contingencies, such as the inaccessibility of a path, by providing the traveller with alternatives at critical points along the way. Of course, this strategy would have to be used sparingly in order to avoid undue complication of the route description. So, although all of the above are genuine possibilities, it is plain that most route descriptions are fairly rigid.

This drawback is linked to another disadvantage of route descriptions. They are highly vulnerable to the effects of various kinds of *noise*. Consider all of the following ways in which we can go wrong when we are trying to follow a route. The route description itself could have been corrupted, either on the page due to a printing error, or in memory if I am not consulting the guide book at every possible opportunity. The landmarks which the route refers to as guidances may no longer exist, or they may no longer be recognisable from the description. Or, I may take a wrong turn because I do not have the background knowledge that the route assumes; I might not, for example, recognise a beech tree if one fell on top of me! On top of all this, even the most diligent attempt to follow the route is susceptible to disruption due to momentary lapses of attention at critical moments. In sum, a typical route description is unlikely to be very robust.

A final significant disadvantage of route descriptions is that they are only *reversible to a limited extent*. Suppose that I want to reach High Beach from my present location in

[33] Travelling in the direction described. The route description may or may not enable us to solve the reverse spatial problems. See the discussion of reversibility, below, for more on this issue.

Upshire; how much help will the route description give me? Part of the reverse route, derived from the original route description in figure 4.2, is shown below:

> From Upshire, a metalled drive goes for 1/2 mile, first twisting left, then winding rightwards before passing a mansion on your right. Pass the iron lodge gates of Woodredon Farm (and the notice that has kept the last mile of your walk free from cars and bicycles); do not miss the pretty Georgian house on the left. The drive leads through the trees away from the farm to the crest of the hill opposite a bus stop (on a road).

The nature of the transformation is easy to understand, as it involves obvious substitutions: "left" for "right", "away from" for "towards", and so on. However, although the new route is closely related to the old one, it is by no means as easy to follow. A moment's reflection will persuade the reader that the main source of this difficulty is the use of distant landmarks as guidances in the original route description: walking away from such a landmark is hardly the same as walking towards it from a particular point. Of course, this does not mean that route descriptions are *never* reversible[34], just that we cannot, in general assume that they are.

Having dwelt at some length on the disadvantages of route descriptions, it is worth mentioning their advantages. Firstly, although they only enable us to solve a relatively small number of spatial problems, such successes are non-accidental; in this sense, we can rely on a good route description to get us to our destination. Secondly, route descriptions tend to be relatively efficient; the route from High Beach to Upshire described in the example is quite direct, for instance[35]. Thirdly, route descriptions do not presuppose any specialised knowledge, barring an understanding of the English language. And finally, route descriptions are relatively quick and easy to use. If all we are interested in is a particular destination, following a route description to it is a good solution.

If routes are easy to use, requiring little or no specialised knowledge, the use of maps is an altogether more elaborate affair. In the first place, we need to be familiar with a whole range of cartographic conventions: the meanings of various symbols, the significance of contour lines, and so forth. Learning how to use maps properly is, in this respect, a matter of

[34] It would be an easy matter to reverse a route description given solely in terms of bearings and distances, for example.

[35] Of course, it is not impossible for a route description to be inefficient; just unlikely, given that they are designed with a specific purpose in mind.

becoming fluent in a new visual language. Using a map to find one's way can therefore be a relatively slow business if one is an inexperienced map reader[36].

What the map lacks in speed, however, it makes up for in other respects. Firstly, a map is, as a rule, not designed with any specific route in mind, but rather impartially represents a large number of routes, integrated within a common space. Hence, a map enables us to solve*many* spatial problems. Once we have the map shown in figure 4.2, for example, we are no longer restricted to following the route from High Beach to Upshire. We can, if we like decide on any number of different walks: Loughton to Waltham Abbey; Claypit Hill to Woodredon Farm; and so on. It is a corollary of this point that, even if we decide to stick to our original itinerary, the map gives us a certain degree of *flexibility*, as it represents many routes between any given starting point and destination, each with different characteristics[37].

Both of the above points are encapsulated in the idea that maps have a *high information content*, relative to route descriptions. This is not simply a matter of the *volume* of information stored in a map, it is also a matter of the highly efficient way all of this information is structured. For each new item which finds its way onto the map is automatically related to all of the other locations represented simply in virtue of the geometrical properties of the plane, and the cartographic conventions. A change in a landscape feature forces us to change just one item on a map, whereas every route description in which that feature plays a part will have to be changed. Such changes may be quite drastic if the landmark played a crucial role in the route description.

Because maps integrate a large amount of spatial information so efficiently, someone who relies on one is capable of exhibiting *robustness* in their response to spatial problems; for a map can suffer a large amount of degradation before it becomes useless. If, for example, one landmark appears to be missing, we can orientate ourselves by means of locating any of the other landmarks shown on the map within egocentric space. Likewise, the map itself may be corrupted in various ways - wearing out along the creases, and so forth - without thereby

[36] I am deliberately glossing over one of the main reasons for this: using a map to navigate involves use of the complex imaginative ability to integrate egocentric awareness of one's surroundings with the objective space of the map. The following reasoning is typical: "I've just passed Walton Street on my left, and Berkeley Precinct now lies to my right; according to the map, that is exactly what I would see if I were standing at a certain point on Ecclesall Road; so that's where I must be, and I need to take the next left for the Botanical Gardens". Following a route description is easier because most of this sort of work has already been done for us.

[37] As noted above, such features as scenic beauty, length, and avoidance of steep climbs are all relevant to the choice of a route.

becoming useless. A surprising amount of information can be lost in this way before the effect becomes debilitating.

Finally, it is fairly obvious that, unlike route descriptions, maps represent spatial relations from no particular point of view (none situated on the Earth's surface, at any rate)[38]. This means that, for any given route a map represents, the *reverse* route is automatically represented. Hence, if one has followed a route using a map, one should also be able to reverse the route to arrive back at one's starting point[39].

I summarise all of the above points of comparison and contrast between route descriptions and maps in Table 4.1[40].

The first six entries on the right hand side of the table are what we're interested in. They correspond exactly with the aspects of spatial orientation we identified in section 4.3. This indicates that having a map that one knows how to use is one way of becoming spatially orientated within an environment. It also suggests that the ability of an organism to find its way around may be explained by supposing that it has access to a map-like representation of its environment.

Before we can appreciate the full significance of this claim for our project, we need to say a little bit more about what maps are. This is the topic of section 4.5.2.

[38] Given the analogy between spatial orientation and understanding, this feature of maps - that they offer a schematic, bird's eye view of the terrain they represent - is reminiscent of Neil Cooper's suggestion, in "The Epistemology of Understanding", that understanding involves a kind of epistemic *ascent.* See p208ff of that paper for more details.

[39] It should not be supposed that the original route and the reverse route will be *equally* easy to follow however. That will depend upon such factors as the way the path forks, and the landmarks available to act as guidances. But whatever the *psychological* difficulties associated with travelling along a route in the opposite direction, it should always be logically possible, as long as one has a map. This is not the case for route descriptions.

[40] The table is from Nadel and O' Keefe, *The Hippocampus as a Cognitive Map*, p89. I have modified it slightly.

Table 4.1: Maps and Route Descriptions

	Route	**Map**
Range of Application	Solve relatively few spatial problems	Solutions to many spatial problems
Non-Randomness	Success is non-accidental	Success is non-accidental
Efficiency	Solutions generally efficient	Solutions generally efficient
Flexibility	Rigid; allow little freedom of choice	Flexible; allow much freedom of choice
Robustness	Not robust; easily rendered useless by the effects of noise or distortion	Robust; can suffer much degradation before they become useless
Reversibility	Not easily reversible	Reversible
Motivation	The final stimulus is the goal; the route is built with this in mind	No object or place on the map is *the* goal; the map can serve *many* purposes.
Information Content	Relatively little	Large amount of information, efficiently stored
Speed	Fast, easy to use	Relatively slow and hard to use
Access	No specialised code needed for access	Knowledge of cartographic conventions required for access

4.5.2 Maps as Models

I now turn to the second principal thesis of this section: that maps are models of the terrain they represent. I take it that most readers will already be inclined to endorse this claim; after all, it just seems intuitively right to think of the relationship between a map and the area it represents as akin to that between a toy car, say, and the real car it's a model of. In this section I draw on the material on models at the end of chapter 3 to build upon this basic intuition.

Previously, we decided that M is a model of O, relative to a set of triples [Q_m, Q_o, T] *iff*:

1. In each triple, Q_m is a set of questions about M, and Q_o is a set of questions about O

2. T represents a set of procedures (algorithmic and/or heuristic), by means of which it is possible, though not necessarily easy, to translate any member of Q_o into a member of Q_m, and any answer to a member of Q_m into an answer to the corresponding member of Q_o.

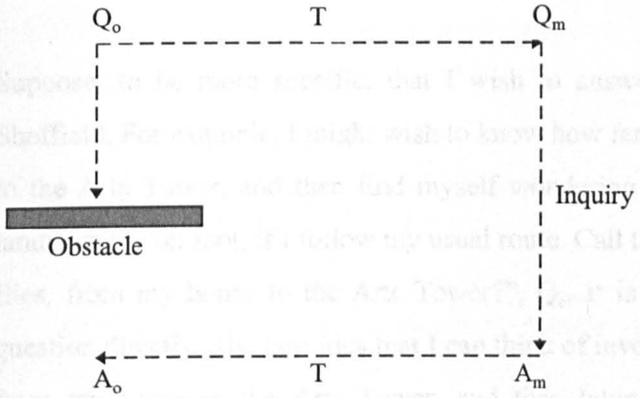
3. T translates correct answers to members of Q_m into correct answers to members of Q_o.

4. M enables us to answer at least one member of Q_o that *could* have been left unanswered during M's construction[41].

Implicit in the above definition is a pragmatic conception of what makes something a model. Models are essentially tools, things that we use to answer certain questions. Typically, we resort to using a model as an *indirect* way of answering a question, when the direct line of inquiry is blocked, or is considered too costly. The following diagram, also from chapter 3, makes this point clear:

[41] As previously explained, this definition is a modified version of that given by Sylvain Bromberger in "Rational Ignorance". In section 3.5, I give his definition and explain why I think it is inadequate. Especially important is the discussion of the difficulty of distinguishing between models and "look-up tables", which condition 4 of my definition is meant to address.

Figure 4.3: The Pragmatic Aspect of a Model



Now let's apply this general framework of ideas about the nature of models to the special case of maps. O, the object of the model is the geographical area represented by the map, and M, the model, is the map itself. Q_o, the set of questions about the object of the model, consists of some of the questions anyone with an interest in the spatial properties of that particular region wants answered. Typically, these include questions of the following form:

1. How far is it, as the crow flies, from A to B?
2. How far is it, on the ground, from A to B?
3. How do I get from A to B?
4. In which direction is A relative to B?
5. What's the shortest route that takes in landmarks A, B, C,... etc?
6. Where's the nearest post office (say) to A?

As the above, by no means comprehensive, list suggests, the set of questions, Q_o, for any given map, is a large and varied one. How large, and how varied depends upon the map concerned. For example, a typical street map of a large city will enable us to answer lots of questions about distances, directions, and routes, whereas the famous map of the London Underground is rather more limited in its application, as it makes no attempt to accurately represent relative distances and directions[42].

It would be a tedious exercise to examine each question on the above list in order to show how the general remarks about the nature of models at the beginning of this section apply to

[42] It's representation of the spatial relations between tube stations is, in fact, principally topological.

it. I shall therefore concentrate on questions about distance, leaving the reader to construct corresponding accounts for the other items on the list.

Suppose, to be more specific, that I wish to answer questions of forms 1 and 2 about Sheffield. For example, I might wish to know how far it is, as the crow flies, from my house to the Arts Tower, and then find myself wondering what the distance between these two landmarks is on foot, if I follow my usual route. Call the question, "How far is it, as the crow flies, from my house to the Arts Tower?", Q_o. It is hard to see how I could answer this question directly. The best idea that I can think of involves hiring a helicopter, flying directly from my house to the Arts Tower, and then taking a reading on a suitable measuring device[43]. But this is too costly in terms of the time, effort, and money involved, so I am forced to adopt the indirect approach of consulting a map of Sheffield.

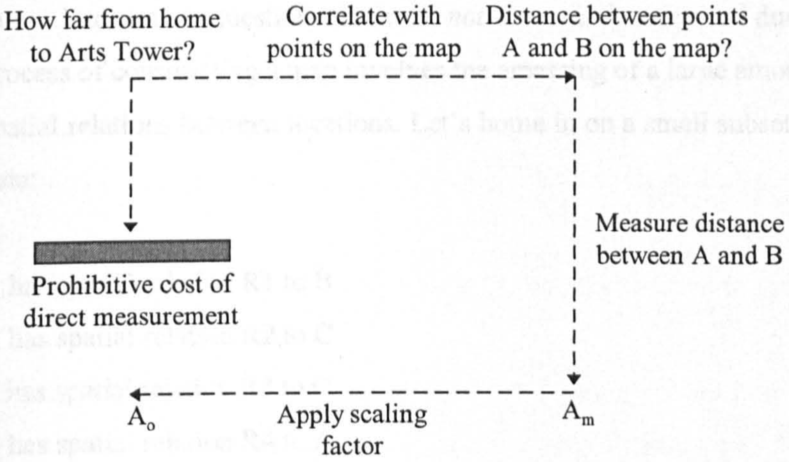
The first step is to translate Q_o, a question about my spatial environment, into Q_m, a question about the map. In this example, I can begin to do this by finding the points on the map that correspond with my house and the Arts Tower. Although my house is not represented*per se* on the map, I know that it is on the corner of Bagshot Street and Psalter Lane, and this is enough to enable me to locate the point that corresponds to it on the page. Call the two points thus identified A (home) and B (Arts Tower). My original question, Q_0, therefore becomes Q_m, "What is the straight line distance between the points A and B?".

Unlike Q_o, Q_m is a very easy question to answer, because of the ease with which the map of Sheffield may be manipulated: all I have to do is measure the distance with a ruler. Call the result of this measurement, which is also the answer to Q_m, A_m. In order to answer my original question, Q_o, I need to apply a suitable translation rule. As far as distances are concerned, the transformation employed by most maps is an extremely simple one. Distances between locations on the ground may be converted into distances between points on the map, and *vice versa*, by applying a constant scaling factor. The scaling factor used in the construction of any given map is usually printed on the cover, in the form "1: 25,000" (say). Let the scaling factor in this example be k; then the answer to my original question about the distance from my house to the Arts Tower is simply $A_o = kA_m$.

[43] It might be possible to consult a satellite or aerial photograph of Sheffield, but such a photograph qualifies as a model on my account. Hence this approach no longer counts as a *direct* attempt to answer the original question. This example also shows that models need not employ arbitrary conventions, although many, including maps, do.

This example shows that using a map of Sheffield to answer a question about the distance between two landmarks conforms to the general schema that presents models as indirect means of answering questions that are too difficult to answer directly. The diagram below should make this point clear:

Figure 4.5: Maps as Models of the Terrain they Represent


How far from home to Arts Tower?
Correlate with points on the map
Distance between points A and B on the map?
Prohibitive cost of direct measurement
Measure distance between A and B
A_o
Apply scaling factor
A_m

What about our second question, the one about the distance on foot between my house and the Arts Tower, if I follow my usual route? In this case, it is still relatively easy to translate the question about the terrain into the corresponding question about the map. The details of this transformation are not important, so I shall leave them out here. The interesting thing about this example is that, having got this far, it is quite hard to see how to answer the question about the map. It is no longer simply a matter of using a ruler to measure the length of a straight line since my route from home to Arts Tower, as represented on the map, isn't straight. Probably the best way of answering a question like this is to take a length of string and shape it along the lines representing the route I follow, measuring the total length with a ruler when I'm done, but working this out takes a fair amount of ingenuity. Hence, this example provides an interesting contrast with the previous one, and illustrates the important point that, although answering a question about the model may be easier than answering the same question about the object of the model, it may still take intellectual effort in the form of creativity, intelligence, and so forth.

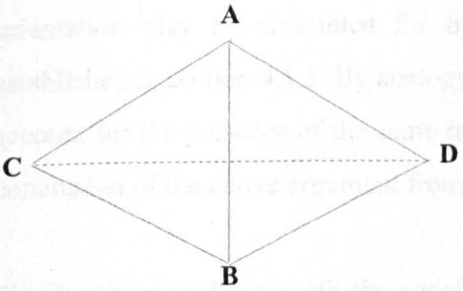
One final point needs clearing up. Remember that one of the main problems with Bromberger's definition of a model was that it failed to mark the intuitive distinction between models on the one hand, and look-up tables or oracles on the other. I introduced condition 4 into the modified definition of a model to deal with this difficulty. But since we haven't yet examined whether maps satisfy this condition, the possibility remains that maps are look-up tables or oracles, rather than models. Although maps *look* like models, we still need to apply the test, since I am using the term "model" in a technical sense that does not necessarily map exactly onto our pre-theoretical classifications.

Condition 4 stipulates that, in order for something to qualify as a model, it must be possible to use it to answer questions that were *not necessarily* answered during its construction. The process of constructing a map involves the amassing of a large amount of data describing the spatial relations between locations. Let's home in on a small subset of this vast collection of data:

A has spatial relation R1 to B
A has spatial relation R2 to C
B has spatial relation R3 to C
D has spatial relation R4 to A
D has spatial relation R5 to B

This information is enough to fix the relative locations of A, B, C and D, and may be used to construct a kind of miniature map, which eventually will form part of a much larger whole:

Figure 4.5: A Fragment of a Map



In the above diagram, the solid lines represent information gathered during the construction of the map. Because of the way the map was constructed[44], the dashed line joining C and D also represents the spatial relation that obtains between those two locations in reality. We can therefore use the map to, for example, find out how far D is from C, *even though this*

[44] By the method of *triangulation*.

information was not explicitly built into it. Of course, we*could* have measured the distance between C and D during the original survey, but the important point is that this would have been unnecessary given the information we had already acquired, since this was sufficient to determine the relative locations of A, B, C and D. In short, the question "What is the distance between C and D?" *could* have been left unanswered (whether it *was* is beside the point) during the construction of the map. But this is precisely what the fourth condition of our definition of a model requires; so, since a map meets all of the other conditions, we may conclude that a map is a model of the terrain it represents.

4.6 Summary

In summarising, let us return to the argument from analogy with which we began section 4.5

P1. Spatial orientation and understanding have many features in common.

P2. In the case of spatial orientation, we can account for the presence of these features by supposing that the organism has a map (usually a mental one) of its environment.

- *Therefore*, it is likely that the presence of the same features in cases of understanding may be accounted for by the existence of a map like representation of the domain in question.

Premise 1, that there are striking similarities between spatial orientation and understanding was established in sections 4.3 and 4.4. Premise 2, that the relevant aspects of spatial orientation may be accounted for by the presence of a map-like representation, was established in section 4.5.1. By analogy, a similar kind of representation may be supposed to account for the presence of the same features in cases of understanding. This establishes the conclusion of the above argument from analogy.

Finally, this, combined with the conclusion of section 4.5.2, that a map is a model of the terrain it represents, enables us to take the final step towards the conclusion we have been aiming for all along: that understanding a given domain is a matter of having an adequate model of it.

Of course, as is the case with any argument that relies upon analogy, this one is not conclusive. But I believe that the resemblance between spatial orientation and understanding is strong enough to make the line of thought traced in this chapter suggestive, at the very least.

In the final chapter of the thesis, I argue that the ease with which the claim that we understand a given domain to the extent that we have an adequate model of it fits in with current thinking about the nature of scientific theories and scientific explanation constitutes further evidence in its favour.

Chapter 5: Scientific Theories and Models

5.1 Introduction

In chapters 3 and 4, I argued that understanding a given domain is a matter of having a model of it, the model providing a kind of intellectual orientation within the domain it represents. This chapter rounds off the discussion of understanding inspired by the *Meno* dialogue by showing how the ideas developed so far may be applied to the important case of *scientific* understanding.

It is reasonable to suppose that the principal aim of scientific inquiry, within a given domain, is usually, if not always, to arrive at an adequate theory of that domain. If the *Meno* is correct in assuming that the goal of inquiry in general is understanding, this implies that scientific theories are sources of scientific understanding - hardly a controversial claim. And if it is true, as I have argued, that understanding a given domain is a matter of having an adequate model of it, this, in turn, implies that scientific theories ought to be models of the domains they represent, even if they do not look much like models, as usually conceived. Sections 5.2 and 5.3 of this chapter aim to persuade the reader that, despite appearances to the contrary, this is indeed the case. It turns out that it is helpful to split the analysis of the claim that scientific theories are models into two parts, one dealing with what I call *microstructure* (section 5.2), and the other dealing with *macrostructure* (section 5.3). The reader will have to bear with me here, as I cannot explain what these terms mean without more background information than I can present in an introductory section.

The idea of scientific understanding is often associated with the concept of scientific explanation, on the plausible assumption that the objective of the latter is to produce the former. This being the case, we should also expect the idea that understanding a given domain is a matter of having a model of it to be relevant to the current debate about the nature of scientific explanation. In section 5.4, I argue that this expectation is met, by showing how the idea of a model may be introduced into each of the three main competing accounts of scientific explanation; doing this illuminates key features of the three accounts, and suggests that they need not be viewed as incompatible after all.

5.2 Theories as Models (I): Microstructure

In this section, I begin to set out the case for viewing scientific theories as models, drawing on the work of such philosophers as Suppes, Suppe, van Fraassen and Giere. All of these philosophers subscribe to what has come to be known as the *semantic* conception of scientific theories. This relatively new way of looking at theories is best seen as a reaction against the *syntactic* conception, developed by the logical empiricists, which was once so widely endorsed that it came to be known as "the received view". Hence, before turning to the details of the semantic conception of scientific theories, I describe the key features of the received view, or at least enough of them to enable the reader to appreciate how it differs from the semantic conception. And, before doing *that*, I say a little about the threefold distinction between syntax, semantics and pragmatics that lies behind this debate. Finally, a word of caution: the term "model" is used in a number of different ways in philosophical discussions of the nature of scientific theories; so, in this section, and the next, we shall need to be clear about how, on any given occasion, the term is being used.

The distinction between syntax, semantics and pragmatics, first articulated by Charles Morris[1], was originally applied to the study of natural language. Since scientific theories resemble languages in having a *representational* aspect, something like this threefold distinction should also apply to them. Hence, in the following paragraphs, I briefly describe the application of Morris' classification to both natural languages and scientific theories.

Syntactic questions concern the relations between the terms of the language, independently of how those terms represent the world or are used by the speakers of the language. Into this category fall questions about the grammaticality of combinations of expressions, and questions about the syntactical rules for transforming one expression into another. Syntactical questions about scientific theories concern what might loosely be described as the internal structure of the theory, the way in which its elements - whatever they may turn out to be - are related to one another. For example, the question of the*logical consistency* of a scientific theory has often been regarded as a syntactical one[2].

[1] In "Foundations of the Theory of Signs", in O. Neurath, R. Carnap, and C. Morris (eds.), *Foundations of the Unity of Science: Towards an International Encyclopaedia of Unified Science*.

[2] Though this is not necessarily the case, since it is possible to define consistency in model-theoretic terms: a sentence, or set of sentences, is logically consistent if it has at least one model, where the word "model" is here used in the *logical* sense described later in this section.

Semantic questions concern the relations between expressions in the language, and the world. Into this category fall many questions which have been central to the philosophy of language this century; questions about the references of names, the extensions of predicates, and the truth of propositions. Within the philosophy of science, there is at least one key debate that concerns semantic questions, namely the one between realists and instrumentalists: does accepting a scientific theory commit us to the view that it states truths about unobservable entities, or does acceptance imply no more than a commitment to its empirical adequacy[3]?

Pragmatic questions concern the relation between linguistic entities and the individuals that use them to communicate. One way of seeing the need for this kind of analysis is to consider the role that indexical terms such as "I", "here", and "now" play in the language. It is impossible to describe the truth conditions of sentences containing these expressions without invoking such factors as the speaker and the context in which the utterance was made. For example, the proposition expressed by uttering a token of the sentence, "I mowed the lawn today", is true if and only if the person speaking mowed the lawn on the day of the utterance[4]. Generally speaking pragmatics investigates the effect of context on what is communicated when language is used. Note that the idea of context is, typically, a broad one: it includes not only the physical environment in which the utterance is made, but also the conversation leading up to it, and the background knowledge, needs and interests of both speaker and audience. As we shall see in section 5.4, pragmatic ideas may be used to shed light on the nature of scientific explanation, one of the principal uses to which scientific theories are put.

We are now in a position to understand the syntactic conception of scientific theories, formulated and developed earlier this century by the logical empiricists. As is well known, these philosophers were much impressed by the advances made in formal logic by Frege, Russell, and others. Since many of them were also working scientists, with an interest in the philosophical foundations of their disciplines, the idea of applying the new insights of formal logic to those issues came naturally. One of the fruits of the resulting intellectual effort was a

[3] The term coined by van Fraassen in *The Scientific Image*; see especially chapter 3 where the concept of empirical adequacy is treated within the semantic conception of scientific theories that he favours.
[4] Some philosophers argue that this feature of indexical terms may be accommodated within a sufficiently broad semantic theory, one which accepts that the sense of a term can be context-sensitive. For further details of this Fregean strategy, see, G. Evans, "Understanding Demonstratives", in Yourgrau (ed.). For the non-Fregean perspective, see J. Perry, "Frege on Demonstratives" in the same volume, and D. Kaplan, "Demonstratives", in Almog *et. al.* (eds.).

view of the nature of scientific theories that was almost universally accepted until the late 1960s. The so called "received view" is a broad church; Carnap, Reichenbach, Hempel, and other philosophers working within the tradition all had their own versions of it. Rather than explore the details of any particular theory, I propose to concentrate on the general features of the received view, using a somewhat crude formulation of the syntactic conception of scientific theories. No intelligent advocate of the syntactic conception would endorse this formulation, but this does not matter, since they all accept one of the key premises upon which it is based: that scientific theories are essentially linguistic entities. As we shall see, advocates of the semantic conception of scientific theories reject this claim.

According to this crude version of the syntactic approach[5], a scientific theory is conceived of as an axiomatic theory formulated in mathematical logic, L, meeting the following conditions:

1. The theory is formulated in first-order mathematical logic with equality, L.

2. The non-logical terms or constants of L are divided into two disjoint classes called vocabularies: (i) the *observational* vocabulary, V_o, containing observational terms; (ii) the *theoretical* vocabulary, V_t, containing theoretical terms.

3. The terms in V_o are interpreted as referring to directly observable physical objects, or directly observable attributes of physical objects.

4. There is a set of theoretical postulates, T, whose only non-logical terms are from V_t.

5. The terms in V_t are given an *explicit definition* in terms of V_o by means of *correspondence rules*, C. That is, for every term "F" in V_t, there is a definition of the form $(x)(Fx \equiv Ox)$ where "Ox" is an expression of L containing symbols only from V_o and logical terms or constants.

The above account of the nature of scientific theories faces many problems. For example, is first order logic really up to the task of representing the claims that scientific theories make

[5] Taken from Frederick Suppe's introduction to *The Structure of Scientific Theories*, p16. In the hundred or so pages that follow he describes a number of refinements to this crude picture, and the fatal problems that even sophisticated versions of the received view face. The interested reader may consult these pages for further references.

about natural laws, or dispositional properties? And how, exactly, are terms in the theoretical vocabulary related to terms in the observational vocabulary; is it really possible to explicitly define the former in terms of the latter by means of correspondence rules? These questions were vigorously addressed by Carnap, Hempel and others, in their efforts to formulate a robust version of the received view, but one central assumption of the syntactic conception of scientific theories was left unexamined: in identifying theories with axiom systems, it is committed to treating them as essentially *linguistic* entities. This is *prima facie* implausible, for at least two reasons:

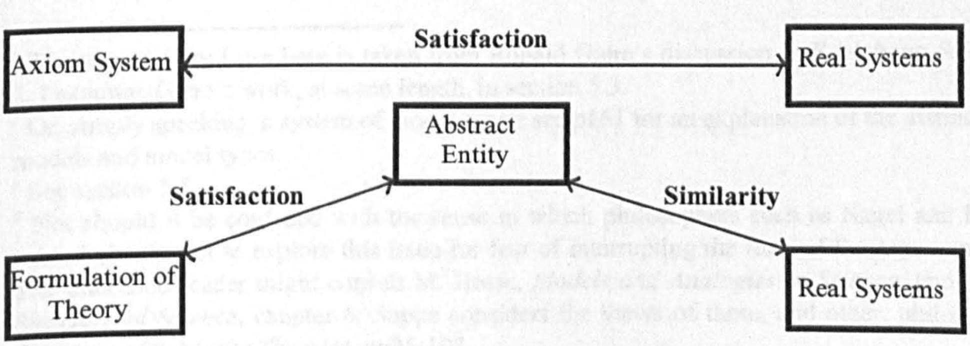
1. Formulations of a theory in two different languages - Maxwell's theory, in English and French, for example - are not two different theories, but two descriptions of the same theory.

2. A single theory can sometimes be formulated in different linguistic terms in the same language; for example, Heisenberg's matrix mechanics, and Schrödinger's wave mechanics are equivalent formulations of quantum mechanics.

Although these are not knock-down arguments - some appeal to the notion of synonymy would at least begin to address them - the underlying worry, that the syntactic conception mistakes the linguistic presentation of a scientific theory with the theory itself, is hard to shake off. Advocates of the semantic conception take this as a sign that the truth about the nature of scientific theories lies elsewhere.

According to the semantic conception, a theory is an *abstract*, as opposed to a purely linguistic, entity. The difference between this and the received view is best appreciated with the aid of a diagram (figure 5.1):

Figure 5.1: The Syntactic and Semantic Conceptions of Scientific Theories



According to the syntactic conception, represented by the top half of figure 5.1, the link between the linguistic formulation of a theory and the real world is direct. In other words, it is assumed that real physical systems falling within the intended scope of the theory satisfy its axioms, insofar as it is possible to deduce observational consequences from those axioms plus the correspondence rules. In contrast, the semantic conception, represented by the bottom half of figure 5.1, insists that the link between the linguistic formulation of a scientific theory and the real systems within its scope is mediated by an abstract entity. *This abstract entity is identified with the theory*, and is simply *defined* as whatever satisfies the axioms, however they are formulated. Hence, according to the semantic conception, the entire *empirical* content of a scientific theory is built into the implicit claim that this abstract entity *resembles* a given range of real physical systems in certain respects[6].

Of course, saying that a theory is an abstract, as opposed to a linguistic, entity does not get us very far, unless we specify the sort of entity we have in mind. In the remainder of this section, and section 5.3, I aim, by drawing upon the work of leading advocates of the semantic conception of scientific theories, to persuade the reader that *a theory is a system of models of the domain it represents*[7]. The term "model", in this context, is used in the sense introduced at the end of chapter 3[8], and should not be confused with the mathematical or logical sense of the term, according to which a model of a formal theory is simply any structure which satisfies the axioms of that theory[9]. To appreciate the distinction, consider the following set of axioms[10]:

A1: For any two lines, there is at most one point that lies on both.

A2: For any two points, there is exactly one line that lies on both.

A3: On every line, there lie at least two points.

A4: There are only finitely many points.

[6] The terminology I use here is taken from Ronald Giere's discussion in *Explaining Science*, chapter 3. I examine Giere's work, at some length, in section 5.3.

[7] Or, strictly speaking, a system of model *types*; see p161 for an explanation of the distinction between models and model types.

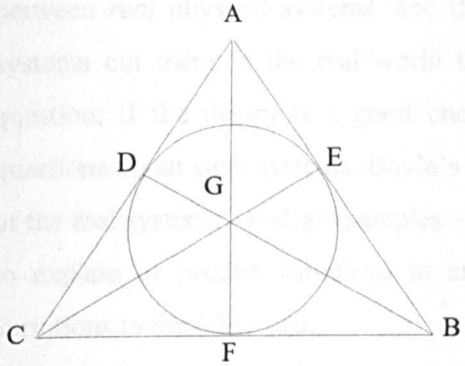
[8] See section 3.5.

[9] Nor should it be confused with the sense in which philosophers such as Nagel and Hesse use the term. I choose not to explore this issue for fear of interrupting the flow of the argument still further. The interested reader might consult M. Hesse, *Models and Analogies in Science*, and E. Nagel, *The Structure of Science*, chapter 6. Suppe considers the views of these, and other, philosophers in *The Structure of Scientific Theories*, pp95-102.

[10] The example is from van Fraassen, *The Scientific Image*, pp41-44.

The above axioms may be taken to define an abstract geometrical structure[11] in the plane, known as *seven point geometry*. This abstract structure is a model, in the logical sense, of the formal theory that consists of the axioms A1 to A4. If the reader finds this abstract structure a little hard to imagine, figure 5.2 may help, since, as I explain shortly, it is a model,*in the sense introduced at the end of chapter 3*, of seven point geometry.

Figure 4.2: A Model of Seven Point Geometry



In this diagram, A, B, C, D, E, F and G represent the seven points, though, strictly speaking, they are not themselves dimensionless. Likewise, AB, BC, CA, AF, BD, CE, and the inscribed circle DEF represent the lines of seven point geometry, although, unlike lines in pure geometry, they have finite thickness. Despite the fact that it is an imperfect likeness, we can still use figure 5.2 to answer questions about seven point geometry. For example, in seven point geometry, could there be a set of three points that all lie on the same straight line? The diagram tells us that there could, since it contains several such sets: C, F and B, to list but one[12]. Note that this question need not have been answered during the construction of figure 5.2, and probably was not answered, given that the figure only needs to satisfy axioms A1 to A4. In short, figure 5.2 satisfies the definition of a model given at the end of chapter 3.

So, when I say that a scientific theory is a system of models of the domain it represents, it is the sense of "model" referred to in the previous paragraph that I have in mind; and I am describing the relationship represented by the *right* arm of the bottom half of figure 5.1. Some confusion may arise because, as we shall see, the *left* arm of the same diagram represents a relationship between the linguistic description of the theory, and the abstract

[11] Or, strictly speaking, a set of abstract structures.

[12] Of course, this is not a question that we should answer simply by inspecting the diagram; some form of geometrical proof is really required to establish the point. I omit this here, for the sake of simplicity. Care is also required in that not all of the questions that can be asked of figure 4.2 are meaningful questions about seven point geometry; for example, questions about the letters of the alphabet associated with particular "points" in the diagram fall into this category.

entity (the theory itself), that is *analogous* to satisfaction. So, curiously enough, something like the *logical* sense of the word, "model" plays a minor role in the semantic conception of scientific theories as well.

Having clarified the sense of "model" intended, we can, at last, begin to set out the case for regarding a scientific theory as a system of models of the domain it represents[13]. The best place to start is with the distinction the semantic conception of scientific theories draws between *real* physical systems, and their *idealisations*. Real physical systems are simply systems out there in the real world that fall within the intended scope of the theory in question; if the theory is a good one, it should enable us to answer a wide variety of questions about such systems. Boyle's law[14], for example, tells us a lot about the behaviour of the real systems - real gas samples – that fall within its intended scope, since it enables us to explain or predict variations in any one of pressure, volume or temperature, given variations in the other two.

Scientific theories enable us to answer questions about the real physical systems they represent by providing us with a set of idealised systems. As the name suggests, these are idealised replicas of real systems; they tell us how the real systems *would* behave under certain specified conditions, that, in fact, do not obtain. Nevertheless, conditions in the real world are close enough to the specified conditions to ensure that the behaviour of the idealised systems approximates to the behaviour of the real systems they represent. To return to our example, Boyle's law employs the concept of an ideal gas, in which collisions between molecules are perfectly elastic, and intermolecular forces are zero. Of course, no real gas sample satisfies these conditions, but many real gas samples nearly do. Consequently, the behaviour of a real gas approximates to the behaviour of an ideal gas over a wide range of pressures, volumes and temperatures.

Let's look at another example in more detail. Consider the behaviour of a simple linear oscillator, a mass suspended between two springs and set in motion, say. As long as we make certain assumptions - that air resistance and friction are negligible, that the springs are perfectly elastic, and so forth - it is possible to show that such a system has the following equations of motion:

[13] In this, and the following paragraphs, I draw on the relevant parts of Frederick Suppe's discussion of the semantic approach in *The Structure of Scientific Theories*, pp221-30.

[14] In case the reader is not familiar with the law, it states that, for a given sample of gas, PV = RT, where P is the pressure of the sample, V its volume, and T its temperature, and R is a constant.

$x(t) = A \cos(wt)$ $v(t) = -Aw \sin(wt)$

Where A is the initial displacement of the body, x(t) is its displacement from its equilibrium position at time t, v(t) is its velocity at time t, and w is a constant which depends upon the elasticity of the springs and the mass of the body.

Of course, no *real* oscillating body obeys these laws, since real systems *are* subject to friction, air resistance, and all the other factors not taken into account in the derivation of the equations of motion. However, we can think of the above equations as describing the behaviour of an idealised replica of a real physical system[15], the behaviour of which *does* depend only upon factors taken into account by the theory. Alternatively, the above equations may be seen as describing how a real oscillating body *would* behave if the only factors influencing it were those taken into account by the theory.

Note that, strictly speaking, it is misleading to say that the above equations of motion *describe* the behaviour of the idealised system, since this tends to suggest that the relationship between the equations and the idealised system is an ordinary empirical one. But this is not the case, since the relationship between the idealised system and the idealisations that characterise it (negligible air resistance, perfectly elastic springs, and so forth), from which these equations are derived, is not an ordinary empirical one. These assumptions, made at the start of the derivation, do not *describe* the idealised system but *define* it in the first place. The relationship between the idealisations and the idealised system is, therefore, analogous to the relationship of satisfaction that holds between a formal theory, and any structure that instantiates the axioms of that theory. Put another way, the idealised system is something like a model, in the *logical* sense, of the idealisations that define it. Although this is an interesting quirk in the semantic conception of scientific theories, advocates of the approach do not tend to dwell upon it[16], their attention being focused mainly on the nature of the relationship between an idealised system, and the real physical systems it represents.

The best way to appreciate the nature of *this* relationship is to consider how we might use the equations that describe the behaviour of the idealised system to answer a question about a

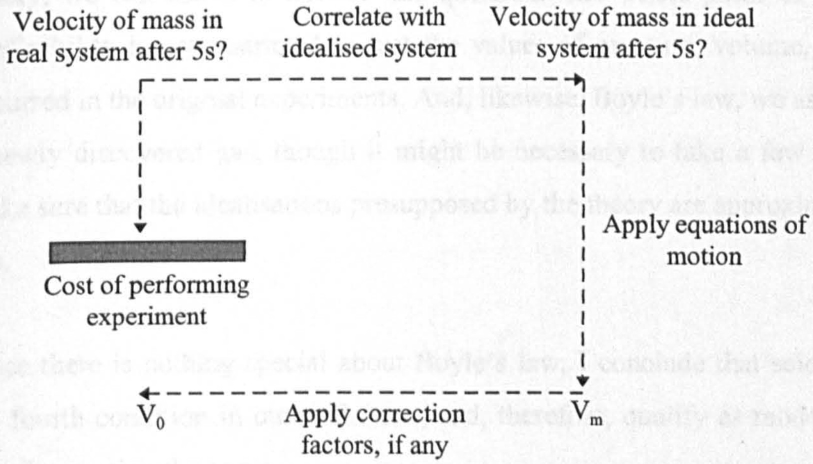
[15] Actually, a *set* of real physical systems - as I explain shortly.

[16] In particular, they tend not to say a great deal about how idealised systems are defined, within the context of a specific theory. This is probably because there is not a great deal, of a systematic nature, *to* say. Such systems may be defined in any number of ways: by means of equations that they must satisfy, statements of ideal conditions that they must meet, and even diagrams that specify the way in which they must behave, to name but a few.

real mass-spring system: "What is the velocity of the mass, five seconds after being released?", say. Call the real system that this question is about O. To answer the question, we need to convert it into a question about the idealised system, which, for convenience,I shall refer to as M. Any theory provides reasonably systematic ways of converting questions about real systems into questions about idealised replicas of them. In this case the question does not change much: "What is the velocity of the idealised point mass in M, five seconds after being released?". The original question is difficult to answer without performing an experiment, but this one isn't: since the behaviour of M satisfies the equations of motion introduced earlier, we can simply plug appropriate values of A and w[17] into the expressions for v(t), to obtain the answer, V_m, say. All that remains is to convert this into an answer, V_o, to the original question about O. Typically, we ignore friction and air resistance in reality, and just assume that $V_o = V_m$, but occasionally the transformation may involve a simple correction factor that takes these other factors into account.

By now, the above procedure should sound quite familiar[18]:

Figure 5.3: The Idealised Mass-Spring System as Model


Velocity of mass in real system after 5s?
Correlate with idealised system
Velocity of mass in ideal system after 5s?
Apply equations of motion
Cost of performing experiment
V_0
Apply correction factors, if any
V_m

As figure 5.3 illustrates, using the idealised mass-spring system to answer a question about a real physical system amounts to using at as a model, in the sense described at the end of chapter 3. It is easy to see that many more questions about real mass-spring systems could be answered in the same way. And, since there is nothing special about the theory of harmonic oscillators, it is reasonable to believe that other theories function in a similar way - as models of the real physical systems they represent.

[17] For the sake of this example, we can take these as given.

However, before we can conclude that theories *are* models, we need to eliminate the possibility that they are look-up tables, as we did when considering the possibility that maps are models in chapter 4. In order to do this, we need to show that theories, *qua* idealised systems, satisfy the fourth condition in the definition of a model, from section 3.6:

> M enables us to answer at least one member of Q_o that *could* have been left unanswered during M's construction.

Fortunately, in the case of theories, it is easy to see how this is likely to be the case. Take Boyle's law, for instance. This simple theory, describing the behaviour of an ideal gas, is the end product of a process of experimentation and reasoning, during which many pressure, volume and temperature measurements were made on real gas samples. Nevertheless, some questions about real gas samples were not answered during the construction of the theory; or at least they need not have been. For example, the question, "What is the pressure of a 1.07243 litre sample of argon, heated to 105.23 degrees Celsius?", could well have been unanswered when the theory was being formulated. But, now that we have an adequate theory, we can use it to answer this question. The whole point of the theory is that its applicability is not restricted to just the values of pressure, volume, and temperature that occurred in the original experiments. And, likewise, Boyle's law, we assume, would apply to a newly discovered gas, though it might be necessary to take a few test measurements to make sure that the idealisations presupposed by the theory are approximately true of the new gas.

Since there is nothing special about Boyle's law, I conclude that scientific theories satisfy the fourth condition in our definition, and, therefore, qualify as models in the sense that I have been using the term.

5.3 Theories as Models (II): Macrostructure

In the previous section, we saw how an idealised system may viewed as a model of a real physical system, and used to answer questions about it. This fact supports the key claim of the semantic conception: that scientific theories are models of the domains they represent. However, in arguing for this conclusion, I deliberately glossed over a number of issues that would have unduly complicated the analysis at the time. Specifically, I wrote, for the most

[18] See chapter 3, p110, and chapter 4, p145.

part, as if a theory consisted of just a single model. But, in general, this is not the case, as closer examination of the two main examples of section 5.2 begins to show.

Taking Boyle's law first, the equation PV = RT may be taken to describe the behaviour of a single model[19], as captured by the concept of an ideal gas[20]. But it sounds a little odd to describe Boyle's law as a *theory* - intuitively, there just doesn't seem to be enough to it; and, in fact, the law is usually seen as part of the much larger body of theory that is classical thermodynamics. A similar point applies to the idealised mass-spring system, which may be taken as a single model[21], but, as such, hardly constitutes a theory. Usually, this type of model is regarded as a small fragment of the much larger body of theory known as classical mechanics.

This suggests that the claim made by advocates of the semantic conception - that a scientific theory is an abstract, as opposed to a linguistic, entity - may be taken in at least two different ways:

1. A scientific theory is a set[22] of models of the domain it represents, sets being viewed as abstract entities.

2. The models that constitute a scientific theory are themselves idealisations which abstract from many features of the real physical systems that fall within its intended scope.

So far, in focusing on a single idealised system and its relation to the real physical system it represents, we have restricted our attention to the second interpretation - to what might be called the *microstructure* of scientific theories. In this section I want to explore the first interpretation in more detail, which concerns what I shall now call the *macrostructure* of scientific theories - the way in which a variety of models are brought together under the auspices of a single comprehensive theory. Throughout the section, I draw on the relatively informal account of the nature of theoretical macrostructure given by Ronald Giere in

[19] Or, strictly speaking, *defines* the behaviour of a single model; see the previous section, p157.

[20] Of course, for the purposes of the theory, there is no need to suppose that there is more than one ideal gas; that is, we can make do with just the one model.

[21] Or, strictly speaking, model *type*. See p161 for an explanation of the distinction between models and model types.

[22] Later, we shall say *system*; see pp168-169.

Explaining Science[23]. Appreciating the way in which a theory typically marshals a whole range of models to represent the real physical systems that fall within its intended scope will enhance our account of the nature of scientific theories, and of understanding in general.

Before we start, we must clear up another ambiguity in the term "model". In section 5.2, we distinguished the logical or mathematical sense of the word, from the sense introduced in section 3.5, derived ultimately from Bromberger's work. Advocates of the semantic conception claim that scientific theories are models, or collections of models, primarily in the latter sense[24]. The new ambiguity occurs, as it were, *within* this sense of the term. To see the nature of the problem, consider the following example:

The motion of a body, of mass m, falling freely in the Earth's gravitational field, is described by the equation $x = gt^2/2$, where x is the distance of the body from the point at which it was released from a state of rest, t is the time elapsed, and g is a constant representing the strength of the Earth's gravitational field. On Mars, where the strength of the gravitational field is h, say, a similar equation of motion, $x = ht^2/2$, holds. Since both equations neglect air resistance and a number of other factors, we must take them to describe the behaviour of idealised physical systems, rather than the motion of a real object falling under the influence of gravity. The question is, though: *how many* models do we have here? It could be argued that there is just *one*, that described by the equation $x = at^2/2$, where a is now a variable representing the strength of the gravitational field, whatever it might be. On the other hand, it also seems right to say that there are *two* models, one representing the motion of a body falling through the Earth's atmosphere, and the other representing the motion of a body falling towards the surface of Mars[25].

The point of this example, is not to force us to choose one answer rather than the other - neither is necessarily wrong - but to encourage us to sharpen up our terminology, in order to

[23] R. N. Giere, *Explaining Science*, chapter 3, especially pp82-7. I have decided not to discuss two other influential accounts within the semantic tradition. Van Fraassen suggests that we view scientific theories as complex structures in state space. The interested reader might consult van Fraassen, "On the Extension of Beth's Semantics of Physical Theories", and *Laws and Symmetries*, pp222-5. Frederick Suppe gives a clear summary of van Fraassen's approach in his introduction to *The Structure of Scientific Theories*, pp226-30. Patrick Suppes' approach is set-theoretical approach, and is developed in a number of papers, including "What is a Scientific Theory?" in S. Morgenbesser (ed.), *Philosophy of Science Today*. Although both of these alternative approaches have some bearing on the idea that scientific theories are sets of models, I believe that Giere's account is more useful for our purposes, since it brings out the connection between models and understanding more clearly.

[24] Though, as we saw towards the end of section 5.2, something like the logical concept of a model plays a small role in the semantic account of scientific theories.

avoid potential misunderstanding. To this end, let us distinguish between the *parameters* of a model, and its *variables*; the best way to appreciate this distinction is to return to our example. In the equation $x = at^2/2$, the strength of the gravitational field, a, is a parameter, since, in using the equation to solve any particular problem, we assign a value to a, and then treat it as a constant for as long as we continue to work on that problem. This, of course, is not to deny that a *can* take on other values, when we work on different problems. In contrast, the displacement of the falling body from the origin, x, and the time elapsed, t, are variables - dependent and independent ones, respectively - since their values change as the idealised system evolves (trivially so, in the case of time). So, the question posed at the end of the previous paragraph comes to this: should we say that two idealised physical systems that differ only in the values of their parameters count as two different models? My view is that it does not matter what we say, as long as we don't get confused; but since we have to make a decision, I propose to restrict the term *model* to idealised systems for which all parameter values are specified. And I shall refer to an idealised system for which at least one parameter value is left unspecified as a *model type*[26,27]. Thus, in our example, there are two different models, both of which are instances of the same model type.

For the sake of clarity, consider another example. We saw earlier that, according to Boyle's law, a simple relationship holds between the pressure, volume, and temperature of an ideal gas[28]: PV = RT, where R is a constant. The interesting thing about this example is that Boyle's law contains no parameters: P, V and T are all variables, since any one of these quantities may change, in a given problem, and R is a physical constant. Therefore, according to the resolution of the previous paragraph, we are dealing with a single model here, one that may be used to predict or explain the behaviour of any real gas sample that falls within the intended scope of the law. Changing the values of P, V and T does not create a different model, but rather defines a different state of the same - the only - model[29].

In sum, we must distinguish between the parameters and the variables of an idealised system: the former are treated as constants for the duration of any particular problem, whilst

[25] Or any planet where the strength of the gravitational field is g or h. respectively.

[26] Van Fraassen notes what is, in effect, the distinction I have just described in *The Scientific Image*, p44. I have followed him in my terminological choices.

[27] In practice, we shall usually be interested in idealised systems for which *all* relevant parameters are left unspecified.

[28] Idealisations involved include: intermolecular forces are negligible; and collisions between molecules are perfectly elastic.

[29] Of course, if the value of R could change, it would be a parameter, like a in the previous example; but, to the best of my knowledge, this cannot, as a matter of physical law, happen.

the latter are viewed as, at least capable of variation[30]. A model *type* is typically defined by an equation, or set of equations, in which the value of at least one parameter is not specified. Once values are assigned to *all* parameters, a particular instance of a given model type is defined, and we have a model *simpliciter*. And when, in addition, values are assigned to some or all of the variables, we have defined a state, or set of states, of the model concerned.

The distinction between models and model types is important because it helps to clarify the nature of the semantic conception of scientific theories. Standard presentations of scientific theories, as might be found in a university textbook, say, rarely, if ever, specify values for the parameters involved. This suggests strongly that when advocates of the semantic conception claim that scientific theories are collections of models, they are really talking primarily[31] about model *types*, as we have defined them. The best way to confirm this hypothesis is to look, in detail, at the presentation of a specific theory - classical mechanics, say.

In *Explaining Science*, Giere gives us a brief tour of part of classical mechanics, as it might be presented in a textbook aimed at first year undergraduates[32]. After preliminary chapters on relevant mathematical techniques, and key concepts such as position, velocity and energy, a typical text introduces Newton's laws of motion, which serve as the foundation for everything that is to follow. At an early stage, attention focuses on the second law, which relates the force, F, acting upon a body at an instant to the mass of the body, m, and the acceleration it experiences at that instant: $F = m\, d^2x/dt^2$ (acceleration being the second derivative of position, x, with respect to time, t).

As Giere points out[33], Newton's laws of motion are the nearest things to axioms that one is likely to find in a text at this level, yet we never see attempts to deduce further theorems from them, as the syntactic account of scientific theories would lead us to expect. This is just

[30] Two further points are in order. Firstly, the function of a given physical quantity may vary depending upon the theory in which it is embedded: a parameter in one theory may be a variable in another, and vice versa. For example, mass is a parameter in Newtonian mechanics, but is variable in relativistic mechanics because of the relationship between mass and energy, $E = mc^2$. Secondly, there's no need to suppose that the distinction between parameters and variables is absolutely rigid even within the confines of a single theory. For a large class of problems which the theory is typically called upon to solve, the same physical quantities will always be treated in the same way, but this does not preclude the possibility of contrived problems where it is necessary to treat a given parameter as a variable, say.

[31] Of course, a collection of model types is always, at least potentially, a collection of models, since we obtain the latter by assigning values to the parameters of the latter.

[32] R. N. Giere, *Explaining Science*, chapter 3.

as well, for with nothing but these laws at our disposal very little, if anything, of interest *can* be deduced. Instead, later chapters of the textbook are devoted to exploring the consequences of specific assumptions about the form the force function in Newton's second law takes. In the first instance, the text is likely to restrict itself to motion in just one dimension, and to consider such cases as the motion of a body falling freely in a uniform gravitational field, and the motion of a body subject to a linear restoring force (the simple harmonic oscillator).

In terms of our previous discussion, this transition from Newton's second law, to the specific forms the force function it contains may take is quite natural. The initial statement of Newton's laws, the second, especially, defines a highly abstract model type. Potentially, this model type has an exceptionally wide range of application, since all massive bodies are assumed to obey the laws that define it. But, in practice, the model type is hard to apply, because in order to do so, we must go right back to first principles, and work forwards - a difficult, and time consuming procedure. Hence, the introduction by the text of specific forms that the force function in Newton's second law might take. In specifying these forms, the textbook, defines further model types, each, as it were, a *sub*-type[34] of the one defined by Newton's laws of motion. The availability of such relatively concrete model types brings us one step closer to being able to apply Newton's laws of motion to real physical systems. However, this is not the end of the matter, since each sub-type may itself have further sub-types.

Consider, for example, the case of harmonic motion. In the simplest case the force function takes the form $F(x) = -kx$, where x is the body's distance from its starting point, and k is a constant. In other words, at any given time, the body is subject to a linear restoring force, proportional to the body's distance from the origin. Substituting this force function into the expression for Newton's second law gives the differential equation: $m\ d^2x/dt^2 = -kx$. It is a simple enough matter to solve this equation to obtain expressions for the position and velocity of the body at time t:

$$x(t) = A\cos(wt) \qquad v(t) = -Aw\sin(wt)$$

[33] *Ibid.* p66.

[34] I assume that this use of the term "sub-type" is fairly intuitive. Clearly the model types obtained by substituting specific expressions for the force function in Newton's second law are simply special

Where $w^2 = k/m$, and A is a constant, equal to the maximum displacement of the body from the origin, otherwise known as the *amplitude* of the oscillation[35].

Having introduced the concept of a simple harmonic oscillator, the mechanics text may then proceed in at least two different directions. One possibility is to remain, as it were, at the same level of abstraction, but to introduce a further degree of complexity into the form of the force function. Typically, the purpose of such a move is to make the model type more realistic, more like the real physical systems it is meant to represent. For example, real physical systems are subject to friction and air resistance, factors ignored in deriving the equations that define the case of simple harmonic motion; so why not incorporate these factors into the model type, by introducing an additional term into the force function? For the sake of simplicity, it is usually assumed that the damping force attributable to the effects of friction, air resistance, and so forth, is proportional to the velocity of the body[36]. This gives us the most common equation for damped harmonic motion: $d^2x/dt^2 = -kx + b\ dx/dt$. As this is not easy to solve, and the details of the solution are not relevant to our project anyway, I propose to skip further consideration of this case[37].

The move from Newton's second law, $F = m\ d^2x/dt^2$, to the differential equation for simple harmonic motion, $m\ d^2x/dt^2 = -kx$, is a move from a highly abstract general model type to a relatively concrete one that applies to a narrower range of real physical systems. The second way of developing the theory of mechanics from this point is to move to an even more concrete level of description. Rather than worrying about damping forces and other complicating factors, the textbook stays with the case of undamped simple harmonic motion, but introduces a new layer of relatively concrete model types. A model type at this level may be so concrete that, once the values of its parameters are set, it becomes a model, in the strict sense, something that we can use to represent the behaviour of real physical systems. Giere describes two examples of model types at this level, both of which may be found in almost any introductory mechanics text[38]. Firstly, there is the mass-spring system, already encountered in section 5.2:

cases of the law. But these special cases are not yet models, in the sense defined above, as they are likely to contain parameters, the values of which are still unspecified.

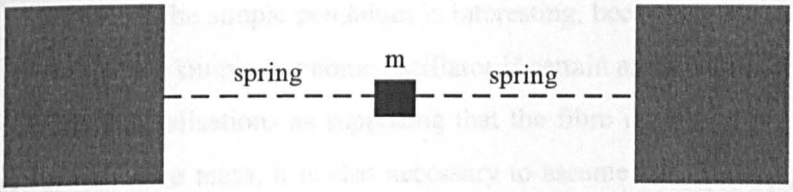
[35] The equations are the same as those that describe the motion of the idealised mass-spring system discussed in section 5.2. This is no coincidence, for, as we shall see shortly, the idealised mass-spring system is a special case of a simple harmonic oscillator.

[36] Further, more realistic, assumptions may be made later, of course. Their effect is usually to complicate the differential equation, and its solution, still further.

[37] For details, see: http://mathworld.wolfram.com/DampedSimpleHarmonicMotion.html

[38] See R. Giere, *Explaining Science*, pp68-71.

Figure 5.4: The Mass-Spring System



In the system represented by the above diagram, the mass m is subject to an initial displacement, x = A at time, t = 0, and then released. This creates tension in the spring, and hence a restoring force which, by Hooke's law is proportional to the displacement from the origin. The constant of proportionality, k which features in the equation of motion, m $d^2x/dt^2 = -kx$, is often referred to as the *modulus of elasticity*, and is a measure of the "stiffness" of the spring. This model type, although relatively concrete, still incorporates a number of idealisations, notably the assumption that the spring is without mass, and obeys Hooke's law perfectly.

The second example of a highly concrete model type that Giere describes is the simple pendulum:

Figure 5.5: The Simple Pendulum

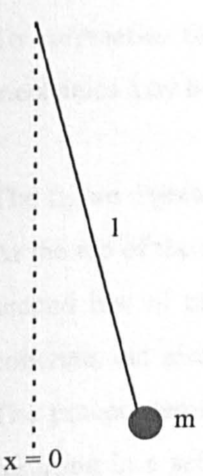


Figure 5.5 represents a body of mass m, attached to the end of a fibre of length l, a short time after being displaced from its equilibrium position and released. After its release, the body is

subject to two forces - the force of gravity, and the tension in the fibre - and oscillates about its equilibrium position.

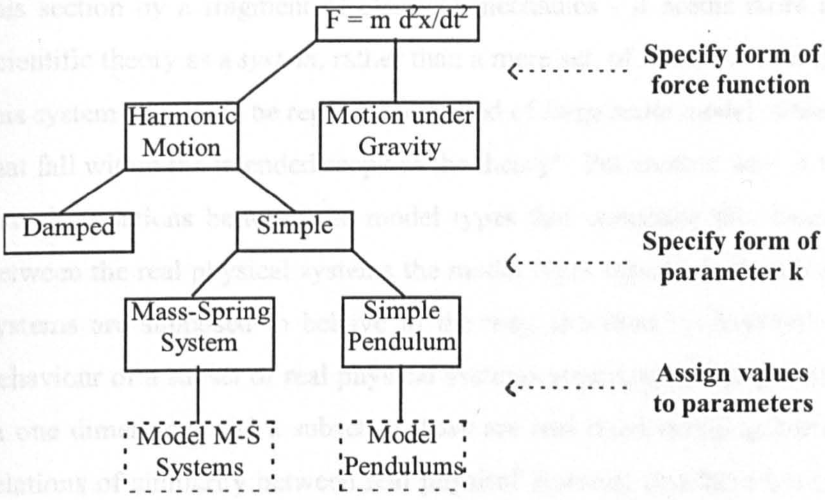
The case of the simple pendulum is interesting, because its equation of motion only reduces to that of the simple harmonic oscillator if certain approximations are made. As well as such common idealisations as supposing that the fibre upon which the point mass is suspended itself has zero mass, it is also necessary to assume that the angle of swing is small. This is because, in tracing out its arc under the influence of gravity, a pendulum actually moves in *two* dimensions. However, if we assume that the angle of swing is small, the vertical component of this motion is negligible, and we may take the pendulum bob to be executing small oscillations in a single dimension. Under these conditions, it is possible to show that the motion of the bob satisfies the equation: $m\ d^2x/dt^2 = -(mg/l)\ x$, where g, as usual, is the strength of the Earth's gravitational field. Obviously, this is just the equation for simple harmonic motion with the value mg/l assigned to the parameter k.

Note that both of the concrete model types described above are derived from the relatively abstract model type of the simple harmonic oscillator by relating the key parameter, k, to real physical quantities that may be determined by experiment. In the mass-spring system, k represents the modulus of elasticity of the spring; and, in case of the simple pendulum, k is a function of the mass of the bob, the length of the fibre, and the parameter, g.

To summarise the discussion so far, according to Giere, a certain fragment of classical mechanics may be represented by means of something like the structure shown in figure 5.6.

The figure represents this particular fragment of classical mechanics as a tree-like structure. At the top of the tree sits the most abstract model type of all, the one described by Newton's second law of motion. As we move down the hierarchy, the model types become more concrete, but also narrower in their range of possible application to real physical systems. The process terminates when values are assigned to all parameters of a given model type, resulting in a set of models capable of representing real physical systems without further mediation.

Figure 5.6: A Fragment of Classical Mechanics



Note: solid boxes represent model types; dashed boxes represent models in the strict sense (model types with values assigned to all of their parameters).

Figure 5.6 also indicates that the key claim of this section, that a theory is a set of model types, may be much weaker than it needs to be. For a set is an *unstructured* collection of entities, whereas, according to the above diagram, theories such as classical mechanics possess a high degree of internal structure. This structure arises as a result of the relationships between the model types that constitute the theory. Relations of *genus* to *species* exist between a model type on one level of the theoretical hierarchy, and the model types on the level immediately beneath it[39]. For example, a mass-spring system is a specific example of the simple harmonic oscillator model type[40], which, in its turn, is a specific example of the highly abstract model type defined by Newton's second law of motion. The hierarchical presentation of the fragment of classical mechanics in figure 5.6 emphasises this kind of relation between model types.

However, theories are also structured by means of relations of *similarity* between model types, that cut across the hierarchy. For example, a damped harmonic oscillator resembles an undamped harmonic oscillator in that both are characterised by a force function containing the term -kx; but it also resembles the case of a falling body subject to air resistance in that both are subject to dissipative forces. For the sake of visual clarity, I have not attempted to represent these similarity relations between model types in figure 5.6.

[39] And also between model types on the "first floor" of the hierarchy, and the concrete models at ground level.

[40] A mass spring system with no damping forces, or other complicating factors, that is.

In view of the complex internal structure that some theories may possess - as illustrated in this section by a fragment of classical mechanics - it seems more accurate to describe a scientific theory as a *system*, rather than a mere set, of related model types. If this is correct, this system may itself be regarded as a kind of *large scale* model of the real physical systems that fall within the intended scope of the theory[41]. Put another way, it is reasonable to expect that the relations between the model types that constitute the theory reflect the relations between the real physical systems the model types represent. For example, *all* real physical systems are supposed to behave in the way specified by Newton's laws of motion; the behaviour of a subset of real physical systems approximates to that of a harmonic oscillator in one dimension; and a subset of *those* are real mass-spring systems. Likewise, there are relations of similarity between real physical systems; and these are represented, within the theory, by similarity relations between the corresponding model types. For example, the motion of a ball bearing falling freely through glycerine resembles the motion of a rusty pendulum, in that both are subject to resistance; and, within the theory, there is a similarity relationship between the corresponding model types (freefall in a viscous medium, and the damped simple pendulum, respectively)[42]. The distinction between macrostructure, the relations between the model types that constitute the theory, and microstructure, the configuration of the model types *qua* idealised representations of real physical systems, is therefore just the distinction between a large scale model and a small scale model of the same domain.

That concludes the presentation of Giere's work on the nature of scientific theories. Although his account is quite informal, compared with those of other advocates of the semantic conception, it is illuminating in the way it draws our attention to what I have called the macrostructure of scientific theories. It also has the advantage of being true both to the way that scientists learn their trade, and to the way that they pursue their trade once their

[41] Figure 5.6 may be regarded as a graphical representation of this large scale model for a fragment of classical mechanics; but, of course, the large scale model itself, like the model types that constitute it, is an abstract entity.

[42] This is deliberately simplistic. *All* real physical systems - apart from those that exist in a complete vacuum - are subject to resistance. In practice, we ignore resistance in many cases, since its effect is negligible. However, this raises the possibility that the same real physical system could be classified differently, depending upon the circumstances. For example, a specific real mass-spring system may be classified as an undamped simple harmonic oscillator, for most purposes; but if greater accuracy is required, the same physical system may be viewed as a *damped* harmonic oscillator. This possibility complicates the mapping of model types onto real physical systems, at least as far as the similarity relationships involved are concerned. However, in practice, most real physical systems *are* treated in a consistent manner by the relevant scientific theories - a ball bearing falling through glycerine will nearly always be treated as an example of freefall through a viscous medium, for example - so the idea that there is some kind of mapping is tenable.

apprenticeship is over. A young scientist *does* cut her teeth on simple problems, for which highly idealised model types suffice, before learning how to solve problems using more realistic model types. And a competent scientist is able to select and apply the appropriate model type from all of those at her disposal to problems about real physical systems.

Before ending this section, it is only fair to acknowledge that the case of classical mechanics is especially favourable to my case, since it is an established theory, with a rich stock of model types to draw upon, and a correspondingly wide range of applications. However, this does not undermine the general conclusion that theories are systems of model types for at least two reasons. Firstly there are other theories of comparably wide scope: classical thermodynamics; classical electromagnetism; and evolutionary biology, to name but a few. Between them these theories cover an extremely wide range of real physical systems, and it is reasonable to suppose that all of them consist of a variety of related model types[43]. And secondly, even a theory that consists of just a single model type may be viewed as a system, albeit a degenerate one, of model types. Personally, I suspect that such theories are quite rare; and, in any case, it is reasonable to suppose that their simplicity is a sign that they are still at a relatively primitive stage of development[44,45].

To summarise the conclusions of this and the previous section:

- *At the level of microstructure*, any reasonably mature scientific theory provides us with a variety of model types, which may be used to solve a wide range of problems about real physical systems falling within the intended scope of the theory.

- *At the level of macrostructure*, the model types that constitute a given scientific theory form a system, structured by vertical genus-species relations, and horizontal similarity

[43] Though, of course, the case for regarding each one as such should, ideally, be worked out in detail.
[44] However, I do not rule out, *a priori*, the possibility of a *bona fide* scientific theory that consists of just a single model type; it's just not the central case, as far as I am concerned.
[45] Another possible counterexample to the claim that scientific theories are systems of model types arises when what looks like a scientific theory consists of just a single *model* (as opposed to model type). For example, plate tectonic theory explains the present distribution of land masses over the surface of the Earth, but this looks like a single model of a unique historical process, rather than a system of model types. My view is that we should bite the bullet in this case: despite appearances to the contrary, plate tectonics is not a scientific theory, but an explanation of a particular process. However, the general principles that this explanation relies upon, assuming there are any, *could* be part of a genuine scientific theory, applicable to a wide range of similar phenomena. For more on the nature of explanation, see the next section of this chapter.

relations between model types. This system may itself be regarded as a kind of large scale model of the theory's domain.

Since scientific theories are prime sources of scientific understanding, the above conclusions support the claim that we understand a given domain to the extent that we have an adequate model of it.

5.4 Models and Scientific Explanation

In this section, I review recent work on the nature of scientific explanation, and show how the ideas about the relationship between models and understanding, developed so far, may be brought to bear upon it[46]. This line of inquiry is a natural one, since it is often claimed that the principal objective of scientific explanation is to produce scientific understanding.

I set the scene with a brief reminder of Hempel's seminal work on scientific explanation[47]. After considering three important objections to it that prefigure more recent accounts, I turn to the current debate[48]. At the time of writing, it is generally accepted that there are three viable approaches to developing a theory of scientific explanation: theoretical unification; causal-mechanical; and pragmatic. In the sub-sections that follow, I describe each of these in turn, concentrating on general features that are distinctive of the approach, rather than on the details of particular theories.

At the end of each sub-section I attempt to show how the ideas about the relationship between models and understanding, developed so far, clarify the nature of the approach to scientific explanation in question, and illuminate its strengths and weaknesses. To be more specific, I aim to show that:

1. A model based approach can handle the difficulties faced by Hempel's, without losing sight of its strengths.

[46] One thing I do not try to do, however, is to develop a theory of scientific explanation based upon the concept of a model; that would be much too ambitious, given the limited space available in this section. Hence, although I occasionally refer to a "model based approach" to scientific explanation in the following pages, this phrase should not be taken to refer to a theory of scientific explanation that I, or anyone else, have actually formulated.

[47] Hempel, C. G. and Oppenheim, P., "Studies in the Logic of Explanation", *Philosophy of Science* 15, pp135-75; reprinted in Hempel's *Aspects of Scientific Explanation*.

[48] As summarised by Wesley Salmon in *Four Decades of Scientific Explanation*; this first appeared as a long article in *Minnesota Studies in the Philosophy of Science* 13.

2. A model based approach is well equipped to explain why the features emphasised by the three competing accounts of scientific explanation are explanatory virtues, since it provides insight into:

- The nature of theoretical unification, and why it promotes understanding
- Why knowledge of underlying causal mechanisms can provide understanding.
- Why good explanation is subject to pragmatic constraints, and what these are.

3. All three contemporary accounts highlight important aspects of scientific explanation, but err in insisting that these represent its essence.

4. It seems possible to reconcile the three supposedly competing accounts of scientific explanation within the single coherent framework provided by a model based approach.

The fact that introducing the concept of a model into the current debate about the nature of scientific explanation makes all of the above possible constitutes further evidence for the claim that we understand a given domain to the extent that we have an adequate model of it.

5.4.1 The Deductive-Nomological Model

Most contemporary philosophical discussions of scientific explanation acknowledge their debt to Hempel's pioneering work. Prior to the publication of his "Studies in the Logic of Scientific Explanation", rigorous analysis of the concept had been one of the more unfortunate casualties of the positivist revolt against the extreme forms of idealism advocated by philosophers influenced by Hegel. It was thought that the link between explanation and understanding made the former suspect, since the latter was regarded as a psychological concept with no place in a philosophical analysis of the logic of science. Although Hempel shared this fundamental concern, he believed that it was possible to rehabilitate the notion of scientific explanation by purging it of its psychological connotations. According to the explication[49] of the concept he proposed, whether X is a scientific explanation of Y is an entirely objective matter, and has nothing to do with the subjective experience of, say, finding Y intelligible in the light of X.

[49] Roughly speaking, explication aims at "philosophically useful concepts that are formulated with precision" [W. Salmon, *Four Decades of Scientific Explanation*, pp5-6]. The explicated concept should resemble the original, but need not capture all of the nuances of ordinary usage if this would compromise its theoretical fecundity. For more on explication, see R. Carnap, *Logical Foundations of Probability*.

As most readers will already be familiar with the details of Hempel's theory, I shall only give a brief summary of it here. Its core is an account of the explanation of particular events or states of affairs within a framework of *deterministic* laws[50,51]. According to Hempel, the explanation of such an event is a deductively valid argument of the following form:

L1, L2... Ln	*Covering Laws*
C1, C2... Cn	*Antecedent Conditions*
E	*Particular Event*

In the argument schema above: E describes the particular event or state of affairs to be explained; L1, L2... Ln describe physical regularities or laws of nature; and C1, C2... Cn describe the conditions that obtained before the occurrence of the event in question. By way of an example, consider the explanation of why my room is so hot on this particular Summer's day. The main reason is that my room is in the attic, and therefore "benefits" from rising hot air from the floors below on warm, sunny days like today. Hence, according to Hempel, the explanation of why my room is so hot today looks something like this:

Hot air rises; Sunlight heats the air...
My room is in the attic; Sun is shining today... etc.
My room is very hot today.

Hempel insists that, for an explanation to be adequate, it must essentially contain at least one general law, and its *explanans* (the statement of covering laws plus antecedent circumstances) must have some empirical content. Roughly speaking then, the idea is that the explanation of a particular event reveals it to be an instance of the operation of universal laws. Given knowledge of those laws, and of the antecedent circumstances, the occurrence of the event could have been predicted with certainty, by means of a deductively valid argument of the above form. For this reason, Hempel's theory is sometimes referred to as the *deductive-nomological* (D-N), or covering law model of scientific explanation.

[50] Hempel later tried to extend the general framework sketched below to cover the case of statistical explanation. His efforts in this direction are generally reckoned to be unsuccessful. For a trenchant diagnosis of the difficulties, see W. Salmon, *Four Decades of Scientific Explanation*, pp68-83, and A. Coffa's article "Hempel's Ambiguity".

[51] Hence, the account that follows does not apply to the explanation of quantum mechanical phenomena, as these are governed, as far as we can tell, by essentially probabilistic laws. Nor does it apply to explanations where the citation of a probabilistic law reflects our ignorance of the relevant deterministic laws. Many of the explanations of applied science - medicine, engineering, and the like - fall into the latter category.

Hempel's attempt to formulate a genuinely objective notion of scientific explanation was initially well received, and enjoyed a remarkably long period of widespread acceptance, for a philosophical theory. However, as time passed, more and more objections to it were raised, and nowadays, it is generally acknowledged to have suffered the death of a thousand cuts[52]. In what remains of this sub-section, I highlight just three of the difficulties that Hempel's theory faces. These have been deliberately chosen to bring out the rationale behind three of the most influential current accounts of the nature of scientific explanation, to be considered later in the section.

The first problem is that Hempel's emphasis on the explanation of*particular* events does not accurately reflect scientific practice, which is geared primarily to the explanation of observed regularities and natural laws. It is natural to assume that it would be easy to extend the D-N model to take account of this fact: why not say that we explain a general law by deducing it from relevant antecedent conditions, together with more general or abstract laws. But this naïve approach does not work, as Hempel realised early on, in the following footnote to "Studies in the Logic of Scientific Explanation":

> Kepler's laws, K, may be conjoined with Boyle's law, B, to form a stronger law K.B; but derivation of K from the latter would not be considered an explanation of the regularities stated in Kepler's laws; rather, it would be viewed as... a pointless "explanation" of Kepler's laws by themselves. The derivation of Kepler's laws from Newton's laws of motion and gravitation, on the other hand, would be recognised as a genuine explanation in terms of... higher-level laws. The problem therefore arises of setting up clear-cut criteria for the distinction of levels of explanation...The establishment of adequate criteria for this purpose is as yet an open problem[53].

This difficulty has prompted some philosophers to develop an account of scientific explanation in which the idea of theoretical *unification* plays a key role. I examine such accounts in section 5.4.2.

The second problem raises doubts about the *sufficiency* of the criteria for successful explanation that Hempel proposes. For many well known examples[54] suggest that something can fit the D-N schema and still be inadequate as an explanation. One kind of counterexample, of which an instance is given below, is especially significant, from our point of view:

[52] A good account of the matter may be found in *Four Decades of Scientific Explanation*, especially Salmon's discussion of the first two decades.

[53] C. Hempel and P. Oppenheim, "Studies in the Logic of Explanation", p273.

Redshift: It is possible to account for the observed Doppler shift in the wavelength of light reaching us from a distant galaxy by supposing that it is receding from the Earth at great velocity. But the shift in the wavelength of light from the distant galaxy does not explain the recession.

This example illustrates the fact that explanations are not generally symmetrical: if X explains Y, Y does not always, or even usually, explain X. The problem is that, if we restrict our attention to the *deductive* relations between the propositions involved, as the D-N model suggests, we find that, in this example, they are symmetrical. In other words, although it is possible, as Hempel might point out, to deduce the value of the redshift given the velocity of the distant galaxy and a background knowledge of electromagnetic theory, it is also possible, given the value of the redshift and the same background knowledge, to deduce the galaxy's velocity. But, intuitively, only the first deduction is explanatory.

So, there are examples in which the deductive relations between explanans and explanandum are symmetrical, even though the explanatory relations are not. Any satisfactory theory of explanation must be able to account for these asymmetries. It is one of the strengths of the causal-mechanical approach, discussed in section 5.4.3, that it does this in a very natural way.

The third, and final, problem I want to discuss is that many *actual* explanations seem to fall well short of the kind of deductive rigour that Hempel describes, without, apparently, being any the worse for it. For an example, we need look no further than the explanation I gave earlier of why my room is so hot on this particular Summer's day. Although I presented this as an example of the application of the D-N schema to a particular case, it is not, as the reader may have noticed, a deductively valid argument; yet there is a strong inclination to accept it as a perfectly satisfactory explanation of the particular state of affairs in question.

So, Hempel seems committed to the counter-intuitive view that, strictly speaking, many (if not all) actual explanations are inadequate, on the grounds that they are not deductively valid arguments, as required by the D-N model. The traditional response of Hempel, and his followers, has been to bite the bullet by distinguishing between the *ideal* explanation of a particular event or state of affairs – the fully fleshed out D-N schema - and an actual explanation of the same event or state of affairs, as given in a specific context. The latter is

[54] For a convenient survey of such cases, see W. Salmon, *ibid.*, pp46-50.

usually a much abbreviated version of the former, but, it is claimed, is only judged to be satisfactory because it approximates to the full D-N explanation. I return to this point in section 5.4.4, where I examine pragmatic theories of scientific explanation.

Before looking at how more recent accounts of the nature of scientific explanation attempt to remedy the above defects, I want to show how the ideas about the relationship between models and understanding developed in this thesis can account for the initial plausibility of Hempel's theory. One of Hempel's key insights is that there is a close relationship between explanation and prediction: the explanation of an event shows that, given knowledge of the antecedent conditions and the relevant covering laws, its occurrence could have been predicted. This is just what we would expect if scientific explanation is a matter of using the model types that a scientific theory provides, for models too are often prized for their predictive value.

To take a relatively concrete example, when designing a new kind of aircraft (e.g. Thunderbird 2!), a scale model is often built first, and its aerodynamical properties studied by means of experiments in a wind tunnel. The purpose of these experiments is to gather valuable predictive information about how the new aircraft will behave under conditions analogous to those in the wind tunnel, without going to the expense of constructing a full scale prototype. Information gathered at this stage will often be used to modify the design of the aircraft, until the model of it behaves in the desired fashion. At this stage, a full scale prototype may be constructed, in the expectation that it too will have the desired aerodynamical properties.

In the above case, the model can also be used to explain the behaviour of the object it represents, in much the same way that it is used for predictive purposes. For example, suppose that we are unwise enough to build a full scale prototype of the new Thunderbird 2 without performing any wind tunnel experiments first. To our dismay, we find that the prototype is not very stable. At this stage we can go back to wind tunnel experiments upon a model to discover *why* the new design is unstable. Such experiments might reveal, for example, that the shape of the jet engines in the new design create turbulence, which destabilises the craft.

If we view scientific theories as systems of model types, as I argued we should in sections 5.2 and 5.3, then we can say similar things about the relationship between their predictive and their explanatory capabilities. For example, classical mechanics provides us with a system of model types for representing the kinetics of a wide range of real physical systems. By assigning values to the parameters of a given model type, we can construct a model[55], an idealised replica, of a real physical system, and use that model to predict - subject to the limitations upon accuracy imposed by the idealisations involved - the behaviour of the real physical system it represents. But, in different circumstances, the very same model may be used to explain some of the properties of the real physical system it represents. To take a concrete example, given suitable equations of motion, we can predict the velocity of a damped pendulum, ten seconds after it is released from rest. But the same equations may also be used to explain why the pendulum has that velocity, ten seconds after being released, or even why it must eventually come to rest.

I conclude that Hempel's theory of scientific explanation has virtues which could be preserved by a theory that employed the concept of a model, but also suffers from a number of defects. Let us now see how more recent accounts of the nature of scientific explanation have tried to remedy these.

5.4.2 Explanation as Theoretical Unification

Like Hempel's account, theories that see a link between explanation and theoretical unification tend to be *deductivist* in character[56]: an explanation is still thought of as, essentially, a deductively valid argument. However, the defenders of such theories[57] part company with Hempel regarding what they take to be the paradigmatic *objects* of explanation. The D-N model is, in the first instance, an account of the explanation of particular events, whereas the approach to scientific explanation discussed in this section sees the explanation of regularities and laws of nature as primary. In this respect, advocates of the view that scientific explanation is closely linked to theoretical unification take seriously the first of the difficulties for Hempel's D-N model that I highlighted a short time ago.

[55] See p161 for an account of the distinction between models and model types.

[56] Though, as we shall see when we look at the relationship between models and theoretical unification, this is not an essential feature of such theories. The point is that, historically, accounts of scientific explanation as theoretical unification have been developed within a broadly Hempelian framework.

[57] These include Michael Friedman ("Explanation and Scientific Understanding"), and Philip Kitcher ("Explanatory Unification and the Causal Structure of the World").

As we saw in the previous section, Hempel did not try to formulate a deductivist account of the explanation of general laws because of the problem caused by the possibility of using an arbitrary conjunction of laws to deduce, and therefore "explain", either of its conjuncts. His reticence proved to be infectious: no one addressed the problem he highlighted for over twenty five years, leaving attempts to develop an account of the explanation of general laws along Hempelian lines in limbo. The silence was finally broken by Michael Friedman, in his paper, "Explanation and Scientific Understanding". In what follows, I attempt to summarise the key features of Friedman's overall approach, as set out in this paper. I do not provide details of his solution to Hempel's problem, since this is rather technical, and not as important, from our point of view, as the special role he assigns to theoretical unification in an account of scientific explanation.

The core of Friedman's theory of scientific explanation is neatly encapsulated in a remark of T. H. Huxley: "In ultimate analysis everything is incomprehensible, and the whole object of science is simply to reduce the fundamental incomprehensibilities to the smallest possible number"[58]. Friedman's version of this claim reads: "... our total picture of nature is simplified via a reduction in the number of phenomena we have to accept as ultimate"[59]. The significance of this statement lies in its implicit denial of the idea that some phenomena are intrinsically more intelligible than others. In making it, Friedman is rejecting a tradition that extends all the way back to Aristotle, who argued that certain things had to be self explanatory if there was to be any scientific understanding at all[60]. The same idea is present in those accounts of scientific explanation that see it as a reduction of the unfamiliar to the familiar[61]. Friedman wants to have nothing to do with any such position; for him, all phenomena stand on an equal footing. As he puts it, intelligibility is not a local property which some phenomena lack until brought into logical contact with other phenomena which have it. Rather it is a *global* property of the corpus of scientific theories we accept at any given time. As we make progress in understanding the world, the number of phenomena we have to accept as ultimate diminishes.

[58] T. H. Huxley, *Darwiniana*, p165; quoted in P. Kitcher, "Explanatory Unification and Causal Structure", p501, n17.

[59] M. Friedman, "Explanation and Scientific Understanding", p8.

[60] Aristotle, *Posterior Analytics*, especially 72b and 99b-100b.

[61] See, for example, P. W. Bridgman, *The Logic of Modern Physics*, ppp37-52, and W. Dray, *Laws and Explanation in History*, Ch. 3. Friedman discusses familiarity theories in "Explanation and Scientific Understanding", pp9-11. Other useful comments may be found in P.Lipton, *Inference to the Best Explanation*, pp27-29.

By way of illustration, Friedman asks us to consider the explanation of Boyle's law in terms of the kinetic theory of gases. How does the fact that Boyle's law may be deduced from certain assumptions about the mechanics of gas molecules make it any more intelligible? Friedman replies that *per se* it doesn't: all it does is replace one brute regularity (Boyle's law) with another (Newton's laws of motion, essentially). However, this is not the end of the matter. The kinetic theory of gases also enables us to derive other empirical regularities such as Graham's law of diffusion, and the values of the specific heat capacities of some gases. It also enables us to see the behaviour of gases as a special case of motion in general, not, in principle, different from the motion of things on the surface of the Earth, or the motion of the planets around the Sun. This is why the kinetic theory of gases is regarded as explanatory - accepting it reduces the total number of independent phenomena we just have to accept[62].

Whatever difficulties might attend the detailed working out of the theory, I suspect that many readers will be sympathetic to the idea behind Friedman's work. Intuitively, new scientific theories that lead to a high degree of theoretical unification are associated with significant advances in our scientific understanding of the world, however difficult it may be, in practice, to say precisely what this amounts to.

Can a model based approach account for the plausibility of linking scientific explanation to theoretical unification? In the remainder of this section, I argue that it can, since:

1. We do not have to see theoretical unification in the deductivist terms in which it is usually couched; it makes just as much sense to view theoretical unification in terms of the merging of large scale theoretical models.

2. A model based approach predicts that the result of merging two or more theoretical macrostructures is likely to increase our understanding of the domains they represent.

3. There are general reasons for regarding unified representations, such as models, as inherently more explanatory than non-unified representations.

[62] Of course, this still leaves Friedman with the problem of *counting* independently acceptable phenomena. essentially this is the problem bequeathed to him by Hempel: given Kepler's laws, K, and Boyle's law, B, why isn't their conjunction K.B a reduction in the sense that Friedman intends? The interested reader should consult M. Friedman, *ibid.*, pp15-19. This attempt to solve Hempel's problem is criticised by W. C. Salmon in *Four Decades of Scientific Explanation*, pp94-101, and P. Kitcher in "Explanatory Unification".

To establish the first claim, we need to recall our discussion of theoretical macrostructure[63], according to which, a scientific theory is a system of model types, the system itself being a kind of large scale model of the theory's domain. To see how it is possible to view theoretical unification in these terms, rather than in the deductivist terms favoured by Friedman, consider the explanation of Boyle's law in terms of the kinetic theory of gases. In terms of our earlier discussion of the macrostructure of scientific theories, this explanation demonstrates that a model[64] that forms part of classical thermodynamics, is a special case of a model type that forms part of classical mechanical theory - roughly speaking, one that involves a large number of point masses experiencing perfectly elastic collisions with one another, and with the walls of their container. The success of the derivation of Boyle's law from the principles of mechanics permits us to view the law as defining a model that may be housed within the macrostructure of classical mechanics.

But, as Friedman points out, the impact upon classical thermodynamics, of the explanatory framework provided by the kinetic theory, is not limited to Boyle's law: Graham's law of diffusion, and the specific heat capacities of gases may all be accounted for in a similar fashion. The net effect of all this explanation is that classical mechanics comes to be seen, at least in principle[65], as capable of assimilating a large part of thermodynamics, hitherto regarded as an entirely distinct theory[66].

So, theoretical unification may be described in terms of the unification of two or more theoretical macrostructures, as characterised in section 5.3. Towards the end of that section, we concluded that this macrostructure, defined by genus-species and similarity relations between the model types that constitute the theory, may itself be thought of as a large scale model of the theory's domain. Hence, theoretical unification, of the kind that Friedman regards as explanatory, may be seen as the merging of two or more large scale theoretical models.

[63] See this chapter, section 5.3.

[64] The reader will recall that Boyle's law defines a model, rather than a model type, since it contains no parameters, and *a fortiori* no parameters to which values have not been assigned.

[65] I add this qualification since, in practice, scientists will continue to use the more convenient model types provided by classical thermodynamics, for a wide variety of problems.

[66] In this respect, the type of case illustrated by this example resembles a hostile takeover bid in the world of corporate finance: classical mechanics expands its scope, at the expense of thermodynamics, which comes to be viewed as a less fundamental theory than was previously supposed. But this type of asymmetry is not a necessary feature of theoretical unification; for example, the unification of the theories of electricity and magnetism in Maxwell's equations looks more like a friendly merger than a hostile takeover of one theory by the other.

To establish the second of the above claims, let us return to the explanation of the properties of gases in terms of the kinetic theory. As noted above the large sale theoretical assimilation this involves means, at least in principle, that classical mechanics inherits the predictive and explanatory power of classical thermodynamics[67], though in practice, scientists will continue to use classical thermodynamics for the sake of convenience. In other words[68], we can use the kinetic theory model type within classical mechanics to solve any problem that can be solved using the model types of classical thermodynamics. But we can also use the kinetic theory model type to solve problems that could not previously be solved by either theory. For example, the derivation of Boyle's law from the kinetic theory of gases may also be used to explain some of the properties of the model defined by the law; in particular, it may be used to account for the value of R, the ideal gas constant[69]. Likewise, we may use the kinetic theory model type to derive an expression for the specific heat capacity of any gas that the kinetic theory may be used to represent.

So, in classical thermodynamics, both the value of the ideal gas constant, and the specific heat capacity of a gas must be accepted as a given, the former being a constant, and the latter a parameter that features in a number of different model types. The above paragraph shows that once we assimilate these model types[70] into the theoretical macrostructure of classical mechanics, this is no longer the case: the value of the ideal gas constant and the specific heat capacity of a gas, are both quantities that can be predicted, or explained, with the aid of the kinetic theory model type. Whenever two or more theoretical macrostructures merge, there are likely to be similar opportunities to explain or predict the values of constants and parameters that had to be taken as given, as long as the theories involved were seen as distinct entities. Since this expands the range of problems we can solve, and thereby increases our understanding of the domains concerned, I conclude that the result of merging two or more theoretical macrostructures is likely to increase our understanding of the domains they represent.

Finally, there are general reasons for believing that unified representations, such as models, promote understanding of the objects they represent *in virtue of their unity*. To see this, recall the discussion at the end of chapter 3[71] about the difference between a genuine model

[67] Or at least the large part of classical thermodynamics that the kinetic theory of gases enables us to assimilate to classical mechanics.
[68] And, again, in principle.
[69] It turns out that $R = N_A k$; where N_A is Avogadro's Number, and k is Boltzmann's Constant.
[70] Or model *simpliciter*, in the case of Boyle's law; see p162.
[71] Chapter 3, pp111-115.

and an oracle or look-up table. At the end of that discussion we added a fourth condition to our definition of a model:

> M enables us to answer at least one member of Q_o that could have been left unanswered during M's construction.

This was the only way we could find of distinguishing a genuine model from an oracle or look up table. The condition reflects the intuition that a model has a kind of unity that an oracle or a look up table does not have, in virtue of which we may put questions to the model that need not have been answered during its construction.

Towards the end of chapter 3, we also noted the likelihood of there being a link between minimal comprehension and look-up tables: someone who never responds to new information with anything more than minimal comprehension, comes to be reliant, in her intellectual life, upon a kind of internalised look up table. It is tempting to see this kind of reliance upon an internalised look-up table as a sign of deficient understanding, and to equate understanding a given domain to having an adequate model of it. The fact that the key difference between a model and a look-up table is that the former has a kind of unity that the latter lacks then suggests that *if*[72] models are sources of understanding, they have this quality in virtue of the fact that they are unified representations of their objects.

If this is correct, the connection between theoretical unification - viewed as the unification of large scale theoretical macrostructures, which are large scale models of the domains they represent - and scientific explanation, is no surprise.

In sum, a model based approach to scientific explanation is likely to be able to preserve the virtues of the theoretical unification approach, whilst accounting for the plausibility of linking scientific explanation to theoretical unification.

[72] I say *if* to avoid charges of question-begging at this point.

5.4.3 Causes, Mechanisms, and Explanation

Causal-mechanical theories of scientific explanation[73] differ from those already discussed in sections 5.4.1 and 5.4.2 in not insisting that explanations are arguments, deductive or otherwise. For, according to such theories, we explain an event or state of affairs by describing the causal mechanisms responsible for its occurrence; and, clearly, there is no need to present this information in argumentative form. Note that the advocates of causal-mechanical theories tend to claim that the citing of causes is the *essence* of scientific explanation, rather than something that merely *contributes* to the explanatory force of a given account. It is obviously possible to accept the latter claim, without endorsing the former, and this is precisely what I intend to do. Like the D-N model, causal theories tend to concentrate, at least in the first instance, upon the explanation of particular events or states of affairs; but their extension to cover the explanation of general laws does not appear to present any fundamental difficulties.

The idea that scientific explanation is, at least partly, a matter of uncovering hidden causal mechanisms tends to lead to the metaphysical picture of a *stratified* physical universe. According to this picture, the universe consists of a number of levels, each with a dual aspect: what happens at a given level is explained by mechanisms operating at the level immediately below it, but may also be used to explain phenomena on the level immediately above it in the physical hierarchy. For example, it is possible to explain many of the observable properties of things around us by describing interactions which go on at the atomic level; but the behaviour of the atoms themselves may, in turn, be explained in terms of the interactions of subatomic particles.

There is little more we need say about the causal-mechanical account of scientific explanation, the details of particular theories of this type not being especially relevant to our project. The basic idea behind the approach is undoubtedly persuasive, with a high degree of intuitive appeal. Think, for example of the urge that many children have to take apart watches and other mechanical devices. Is it not plausible to view this as a manifestation of the desire to understand how things work? And is it not the case that this desire is, at least partly, satisfied when the hidden mechanisms of the object concerned are revealed?

[73] As defended by, among others, Wesley Salmon in *Scientific Explanation and the Causal Structure of the World*, and Peter Railton in "Probability, Explanation, and Information" and other papers.

But, apart from its *prima facie* plausibility, the causal-mechanical approach has other advantages. Chief among these is the ease with which it handles the second difficulty for Hempel's D-N model highlighted in section 5.4.1, the existence of asymmetries of explanation that are not reflected in the corresponding D-N schema. For causal relations too are asymmetrical: if X caused Y, then Y was not the cause of X. And since, according to the causal mechanical account, we explain an event by identifying its causes, the asymmetry of explanation is simply a manifestation of the asymmetry of the underlying causal relations involved. To return to the counterexample to Hempel's account described in section 5.4.1, the recession of a distant galaxy explains the redshift observed in its spectrum, but not *vice versa*, because movement away from the point of observation causes redshift, but not *vice versa*.

The most interesting *problem* with the causal mechanical approach, from our point of view, is that it is not easy to extend it in the direction of an account of the nature of explanation in general[74]. Not all explanations are causal, there being many disciplines where the idea that explanation is a matter of exhibiting hidden mechanisms is inappropriate:

Example 1: Many classes of equations in one variable have solutions which can be expressed as rational functions of their coefficients. This is trivially the case for linear equations, for if $ax + b = 0$, then $x = -b/a$. Most readers will also recognise the expression for the roots of a quadratic equation: if $ax^2 + bx + c = 0$, then $x = [-b \pm (b^2 - 4ac)^{1/2}]/2a$. It is possible to derive similar, though much more complex, expressions for the roots of cubic and quartic equations. Until the end of the eighteenth century no one knew whether analogous results could be obtained for equations in higher powers of x; or if they couldn't, why that should be the case. Then Galois, drawing upon his work in group theory, succeeded in deriving a criterion for the expressibility of roots as rational functions of the coefficients. This general theory explained both the special results already mentioned, and the absence of corresponding results for equations in higher powers of x[75].

[74] Defenders of the causal-mechanical approach tend to disown any interest in providing an account of explanation in general, or, more strongly, claim that the notion is an incoherent one anyway. My view is that it is too early to make such claims, and that we must therefore be circumspect about adopting accounts of scientific explanation that close off potentially fruitful avenues of inquiry.

[75] This example is from P. Kitcher, "Explanatory Unification and the Causal Structure of the World", p425. He gives a number of other interesting examples of mathematical explanation in the section of the paper from which this is taken.

Example 2: Why do some scientific explanations exhibit the kind of asymmetry highlighted in section 5.4.1? Causal-mechanical accounts provide us with an excellent explanation of this fact: the asymmetry of explanation in these cases simply reflects the asymmetry of the underlying causal relations. But this is not itself a causal explanation!

Example 3: I visit friends in nearby towns at weekends, usually travelling by bike. The places I visit lie in many different directions from my house yet, I always end up having to cross a busy dual carriageway. Why is this the case? A map of the city provides the answer: my house is situated in the area bounded by the ring road, whereas all of my friends' houses lie outside it. Hence it is topologically impossible for me to visit a friend without crossing the dual carriage way.

The first two examples are drawn from the fields of mathematics and philosophy respectively. In neither case does the fact to be explained look as if it is amenable to the type of causal explanation favoured by mechanistic accounts. The third example is interesting because it involves a process - a bicycle ride - for which we *would* normally seek a causal explanation; but, in the circumstances, this type of explanation is rejected in favour of what is essentially a topological one[76].

In my view, the above examples show that the causal-mechanical approach to scientific explanation is likely to be unduly restrictive in the long run. However, it does contain an important element of truth that can be appreciated in the light of our earlier discussion of the relationship between scientific theories and models[77]. The best way to see this is to ask what is it, exactly, that the models that make up a scientific theory model? The most plausible answer, I believe, is: the entities falling within the scope of the theory *and the causal relations between them.* If this is correct, then it raises the possibility that the causal mechanical nature of many, if not all, scientific explanations is a feature of them *qua scientific*, rather than *qua* explanation. Given that the objective of scientific inquiry, within a given domain, is to construct adequate models of its entities and the causal relations between them, it is hardly surprising that scientific explanations that presuppose those models will describe the causal relations they represent. This suggestion leaves open the possibility that

[76] P.Kitcher discusses this kind of case in "Explanatory Unification and the Causal Structure of the World", pp422-28. The same section of the paper also contains other examples of explanation in mathematics. For more on mathematical explanation, see M. Steiner, "Mathematical Explanation", and M. D. Resnik and D. Kushner, "Explanation, Independence and Realism in Mathematics".

[77] Sections 5.2 and 5.3 of this chapter.

explanation in other disciplines - philosophy, mathematics, and the like - need not cite hidden causal mechanisms - and a good thing too, given the examples above.

This shows that a model based account of the nature of scientific explanation can co-opt one of the key virtues of the causal-mechanical approach - its adroit handling of the asymmetries of explanation that create problems for Hempel's theory - without succumbing to one of its main defects. A model based approach to scientific explanation may also explain *why* knowledge of underlying causal mechanisms might be explanatory. To see this, we need to recognise that causal explanations facilitate technical control: if I know that A causes B, then I know that I can bring about B by bringing about A. Since we have previously[78] characterised understanding as the ability to solve many problems (and to solve them efficiently, and so forth) relating to a given domain, one of the messages of the thesis is that information which brings things under our control contributes to our understanding[79]. For example, the kind of understanding that a map provides is tied to its usefulness for solving a wide variety of spatial problems, thereby giving us a form of control over our environment. Hence, it is not surprising that knowledge of underlying causal mechanisms, in general, should be regarded as potentially explanatory.

In sum, a model based approach can account for both the strengths and the weaknesses of the causal-mechanical approach to scientific explanation, and can preserve the former without adopting the latter. In addition, a model based approach to scientific explanation leaves open the possibility of a general account of explanation further down the road.

5.4.4 The Pragmatics of Explanation

Pragmatic accounts of explanation take their cue from the third difficulty for Hempel's D-N model identified in section 5.4.1, the fact that many real explanations fall well short of the level of detail that Hempel and his followers would lead us to expect. For example, pointing out that a nearby rock contains uranium would usually be enough to explain the clicking of the Geiger counter in my hand. But, clearly, this is nothing like the full D-N explanation of the event that Hempel envisages. Causal-mechanical accounts of explanation face a similar problem, insofar as they insist that the *full* explanation of a particular event or state of affairs is a complete description of its causal history.

[78] See sections 4.3 and 4.4.

[79] Not that I wish to reduce understanding to the ability to exercise technical control, of course.

As noted in section 5.4.1, the usual response of Hempel, and his followers, to this objection, is to distinguish between the *ideal* explanation of a particular event or state of affairs – the fully fleshed out D-N schema - and the actual explanation of the same event or state of affairs, as influenced by the pragmatic factors operating in that specific context. As the example of the previous paragraph shows, the latter usually contains much less detail than the former; but, according to Hempel, it is only acceptable as an explanation because it approximates (however imperfectly) to the full D-N version. Recently, Peter Railton, a defender of the causal-mechanical approach to explanation, has developed a sophisticated response to the concerns of pragmatists along broadly Hempelian lines. In the remainder of this section, I describe his approach to the issue, using it as a way of clarifying what it is that is distinctive of pragmatic accounts of explanation.

Following Hempel's lead, Railton suggests that, in order to make sense of actual explanatory practice, we need to distinguish between what he calls the *ideal explanatory text* and *explanatory information*. As he advocates what is essentially a causal-mechanical approach to explanation, he believes that the ideal explanatory text should include the complete description of the causal history of the particular event or state of affairs to be explained. In addition, it will include statements of the fundamental principles and laws of nature that lie behind the occurrence of the particular event[80]. Clearly, the ideal explanatory text is liable, in general, to be formidably complex, since it must include details of all relevant antecedent conditions, and of all the causal mechanisms responsible for the occurrence of the event or state of affairs in question[81]. For this reason, exhibition of the ideal explanatory text is rarely, if ever, attempted in practice. Instead, scientists are content to supply one another with explanatory information - information about the ideal text given in response to a specific request for explanation[82]. Explanatory information coincides with a (usually small) portion of the ideal explanatory text which, in the context of inquiry, is viewed as especially salient. In the above example, for instance, pointing out the presence of uranium in the nearby rock supplies a small part of the ideal explanatory text for the Geiger counter's clicking. Of course, the complete text contains much more information: statements of the fundamental principles of atomic physics; a description of the construction and operational principles of

[80] And it may include other things - illustrative diagrams, and the like - besides.

[81] I am deliberately skirting round the details of what Railton considers to be a complete explanation, as the general points I want to make are not affected by them.

[82] Railton's notion of information resembles that of Wiener and Shannon, the founders of communication theory, except that his is a semantic, as opposed to a purely syntactic, conception. See Railton, *ibid.*, pp244-46. For a clear introduction to communication theory, and a semantic theory of information, see F. Dretske, *Knowledge and the Flow of Information*, chapters 1-3.

Geiger counters, and so forth. But, in the circumstances, it is the proximity of uranium that is especially salient; hence it is volunteered as a piece of explanatory information.

At first sight, this looks like a significant concession on the part of the causal-mechanical theorist, since it acknowledges the role played by context and other pragmatic factors in scientific explanation. However, in this case, appearances are deceptive, and the core of the causal-mechanical approach remains intact. To see why this is so, we need to recognise that advocates of the causal-mechanical approach typically believe that it is possible to assign a *realistic* interpretation to their talk of underlying causal mechanisms[83,84]. Therefore, they still have - and usually exercise - the option of claiming that explanatory relations are *objective* features of the physical universe; they obtain whether anyone thinks that they do or not. The distinctive feature of the pragmatic approach to scientific explanation is that it rejects this claim. In its place, it insists that explanation is *always* a matter of citing information that is salient in a given context of inquiry, where what is salient depends upon the needs and interests of the participants in that particular explanatory context.

In effect, the causal-mechanical approach to explanation claims that causal relations between events are explanatory *per se*, whether or not these relations are ever described in actual explanations. The pragmatic approach claims that it is the citing of information in a specific context of inquiry that makes it explanatory. Of course, this aspect of the pragmatic approach is quite consistent with the claim that, as far as *scientific* explanation is concerned, the information cited in response to a request for explanation must identify *causal* relations. The point is simply that without such a context, this is just another piece of descriptive information, with nothing to mark it out as distinctively explanatory. The point is neatly put by van Fraassen, one of the most sophisticated advocates of the pragmatic approach: all previous theories of scientific explanation have failed because they try to treat it as a two place relation between the fact to be explained, and the explanatory theory; but, in reality, we are dealing with a three place relation between fact, theory and *context*[85]

[83] Hume, of course, did not think it was. See *An Enquiry Concerning Human Understanding*, sections IV-VII. Much of Salmon's *Scientific Explanation and the Causal Structure of the World* is devoted to developing a response to Hume's challenge. For a sceptical assessments of this attempt, see B. C. van Fraassen, *The Scientific Image*, pp118-23, and P.Kitcher, "Explanatory Unification and the Causal Structure of the World", pp459-77.

[84] Whether advocates of the causal-mechanical approach *have* to interpret their talk of underlying causal mechanisms realistically is another matter. Their tendency to do so is, perhaps, more a reflection of the mistrust of subjectivity that, historically, has shaped philosophical accounts of scientific explanation, than of the inherent logic of their position.

The aim of the last few paragraphs has been to clarify the way in which pragmatic accounts of scientific explanation differ from the other approaches we have examined in this section. Of course, there remain many issues that need to be addressed by a pragmatic theory of scientific explanation: what, for example, constitutes a satisfactory response to a request for explanation, and how does context restrict the range of acceptable responses to such a request? Specific pragmatic theories of scientific explanation attempt to answer questions like this in different ways[86], but this is not something that we need to worry about, as long as we are clear about the key ideas that all such theories share.

The idea that a scientific theory is a system of model types sheds light on the nature of the pragmatic approach to scientific explanation, and the reasons for it plausibility, in at least two ways. Firstly, as we saw in chapters 3 and 4, a model is a unified representation of a particular domain; as such, the same model may be used to solve a wide range of problems relating to the domain it represents, the salient information, in each case, being determined by the nature of the problem, and a variety of contextual factors. For example, a map of our environment enables us to solve any number of spatial problems relating to it. In particular, the use of a map to direct someone has an essentially pragmatic dimension, since the directions we give depend upon the location of the individual concerned, and upon such factors as whether they have a car or not, whether they are disabled, and so forth.

In a similar fashion, the undamped mass-spring model, described in the previous section, may be used to explain various aspects of the behaviour of real mass-spring systems. For example:

- *Why is the motion of such systems periodic*? Because the equations that describe the motion of idealised replicas of such systems have a regular period, these equations of motion being consequences, ultimately, of Newton‘s second law of motion.

- *Why does this particular mass-spring system have the period that it does*? Because, according to the undamped mass-spring model type, the period of such systems depends upon the modulus of elasticity of the springs, and the mass of the body; and in this particular case, the modulus of elasticity is K, and the mass of the body is M.

[85] B. C. van Fraassen, *ibid*, p156.

[86] For van Fraassen’s answers, see *The Scientific Image*, pp146-51.

Note that in the second example, an adequate response to the original question may simply be to state the mass of the body, or the modulus of elasticity of the springs; it depends upon what we can assume our audience already knows. Likewise, the depth of the explanation given in the first example will depend upon how interested the inquirer is in the mathematics underlying the undamped mass-spring system model type.

The second way in which the idea that a scientific theory is a system of model types sheds light on the pragmatic approach to scientific explanation, is that, depending upon our purposes, different models may be appropriate. Again, the map example illustrates this point most clearly. If I need to get from Oxford Circus to Waterloo station, access to a map of the London Underground is likely to be useful. However, if I have neither a valid tube ticket, nor the money to buy one, a street map of that part of London will be much more useful, since the most sensible plan is to aim to reach Waterloo on foot.

If a scientific theory is a system of model types, some of the pragmatic aspects of scientific explanation may be attributable to the availability of multiple model types, all having some bearing upon the real physical system under consideration. For example, we have already seen how the theory of undamped harmonic motion may be used to account for certain features of a real mass-spring system - the periodic nature of its motion, the length of its period, and so forth. But if, instead, we wish to account for the fact that the system eventually comes to rest, we must use a model type that takes account of the damping forces within the system, and demonstrate that the resulting equations of motion predict that such a steady state will obtain, in the long run[87]. Similarly, if we wish to account for the fact that the equations of motion for simple undamped harmonic motion do not apply to the system[88] if the initial displacement of the mass exceeds a certain critical value, we must use a model that acknowledges the fact that, in real life, springs are not perfectly elastic, and may be stretched to a point where Hooke's law no longer holds for them.

This last example suggests that, in giving explanations that relate to a particular physical system, we are not restricted to model types that fall within the scope of a single theory. For someone may want to know *why* Hooke's law fails to hold beyond a certain point, seeing this as part of the full explanation of why the mass-spring system no longer behaves like a

[87] Use of a model type that takes account of damping forces implicitly acknowledges that answers obtained using the original model type were approximate. But as long as those answers were accurate enough for the purpose for which they were required, this does not matter.

simple harmonic oscillator. We can respond to this request for an explanation, using model types that represent the bonds that exist between molecules in a solid, and how those bonds react when an external force is applied; but these are usually thought of as model types belonging to the theory of the solid state, rather than to classical mechanics[89]. Or, perhaps the spring failed because it was rusty, and someone wants to know why this was the case, again seeing it as an essential part of the full explanation. In this case an explanation that made use of the model types found in basic chemical theory is called for.

It is interesting to note, in this connection, that significant theoretical unification opens up the prospect of the wholesale application of new model types to real physical systems that they would not have been used to represent before. For example, the availability of the kinetic theory of gases means that thermodynamical explanations of the behaviour of real gas samples can be supplemented with explanations based on statistical-mechanical model types, where necessary. And, in some cases, this opens up the possibility of explaining things that once had to be accepted as given, such as the values of the specific heat capacities of some gases.

In sum, pragmatic accounts of scientific explanation claim that there is an *essentially* pragmatic dimension to scientific explanation. This is what a model based approach to scientific explanation would lead us to expect, since:

1. The use of a given model type to solve a range of problems, including its use in formulating explanations, has a pragmatic dimension.
2. The selection, from the range available, of the best model type to solve any given problem has a pragmatic dimension.

In subsections 5.4.2 and 5.4.3 we came to analogous conclusions about the theoretical unification and the causal-mechanical approaches to scientific explanation. A model based approach to scientific explanation promises to be able to account for both the plausibility of

[88] In the sense that they cannot be used to make accurate predictions (accurate enough for the purposes at hand, that is) about the behaviour of the system.

[89] Not that the boundary of a scientific theory need be sharply defined. In practice it may be difficult to assign model types to one theory rather than another; and, in some cases, there may not be a fact of the matter, anyway. The cited example is a case in point, since it is, strictly speaking, a mechanical model type, but is not the kind of thing that would normally be included in a mechanics textbook. One

linking scientific explanation to theoretical unification, and the likelihood that the citing of causes should play a key part in many, if not all, scientific explanations.

In sum, a model based approach to scientific explanation can accommodate, and account for, the key virtues of the other approaches examined in this section. Although these approaches are usually thought of as being in competition with one another, our findings suggest that this may not be the case. The more likely scenario is that all of these approaches latch on to something significant about the nature of scientific explanation, something that is already implicit in the idea that a scientific theory is a system of model types, but then make the mistake of claiming that this feature is the*essence* of scientific explanation. But this last step is not necessary; for the idea that a scientific theory is a system of model types provides us with a way of seeing theoretical unification, the citing of causes, and the influence of pragmatic factors, as aspects of the same phenomenon[90].

5.5 Summary

The aim of this chapter has been to show that contemporary work in the philosophy of science provides further evidence for the claim that we understand a given domain to the extent that we have an adequate model of it.

In sections 5.2 and 5.3, we examined the semantic conception of scientific theories, and concluded that:

- At the level of microstructure, a scientific theory provides us with a variety of model types; these may be used to solve problems about real physical systems falling within the intended scope of the theory.

- At the level of macrostructure, a scientific theory is a *system* of model types, structured by genus-species and similarity relations; this system is itself a kind of large scale model of the theory's domain.

would be much more likely to find discussion of this model type in a chemistry or solid state physics textbook.

[90] I least it holds out the promise of doing so - I do not pretend to have presented a comprehensive, unified account of the nature of scientific explanation in this section; nor was that my aim.

Since scientific theories are prime sources of scientific understanding, these conclusions support the claim that we understand a given domain to the extent that we have an adequate model of it.

In section 5.4, we turned to the nature of scientific explanation, considering three contemporary approaches to the issue: theoretical unification; causal-mechanical; and pragmatic. Our conclusions were:

- In each case, the idea that a scientific theory is a system of model types may be used to shed light on the nature of the approach concerned.
- The idea that a scientific theory is a system of model types helps us to reconcile the three approaches, since it enables us to see them as describing three complementary aspects of a single phenomenon.

Since it is widely acknowledged that there is an intimate relationship between scientific explanation and scientific understanding, the above conclusions provide further evidence for the claim that we understand a given domain to the extent that we have an adequate model of it.

Conclusion

As this has been a long thesis, and has touched upon many disparate topics, a detailed summary of its argument is called for.

First, recall the two claims that we set out to establish[1]:

- The aim of rational inquiry is understanding; this is achieved to the extent that we end up with an adequate model of the object of our inquiry.

- The possibility of rational inquiry presupposes fundamental hopes, which, in the circumstances, it is legitimate for us to entertain.

Chapter 1 began the long process of establishing the above conclusions by examining the way in which the key themes of the thesis - hope and understanding - are treated in Plato's *Meno* dialogue, and Peirce's writings on the nature of scientific inquiry.

In Plato's *Meno*, the idea that understanding is the goal of inquiry emerges in the context of Socrates' dialogue with Meno's slave[2]. At the end of their conversation, Socrates implies that the slave boy does not yet have *episteme*, the ultimate goal of inquiry, even though he has a justified true belief, and there is nothing to suggest that anything more is required for knowledge in the circumstances[3]. This suggested that a better translation of *episteme* might be "understanding", and raised the question of what more the slave boy needs to do to convert the justified true belief he has at the end of the dialogue into understanding. Further examination of the dialogue uncovered two clues regarding the nature of understanding:

- Understanding is, at least partly, a result of trying to make sense of new information by seeing how it squares with prior beliefs and intuitions[4].

- When one has understanding of a given domain, one's beliefs enjoy a distinctive and valuable form of stability[5].

[1] See the introduction to the thesis, p1.

[2] See pp20-23 for further discussion of this dialogue.

[3] That is, there are no unusual Gettier-like complications. See Gettier, "Is Justified True Belief Knowledge?".

[4] See pp9-10.

Further consideration of the nature of understanding was then postponed until chapter 3.

The idea that rational inquiry presupposes certain fundamental hopes emerges, in Plato's dialogue, in the wake of Socrates' solution to Meno's famous paradox[6]. This, it will be recalled, rests upon the Theory of Recollection, the idea that the soul has already experienced all truths in a previous existence, all learning in this life being merely recollection of them. Socrates admits that we cannot *know* that the Theory of Recollection is true, but we do have a reason for believing it, because doing so will make us more vigorous inquirers, and will insulate us from a kind of epistemic despair. The justification of the Theory of Recollection is therefore a pragmatic one, and our attitude towards it, if we accept it, begins to look more like one of hope than belief.

Consideration of Peirce's work on the nature of scientific inquiry reinforced this suspicion[7]. As we saw, towards the end of chapter 1, one of the distinctive features of Peirce's theory of inquiry is the amount of time he devotes to the problem of *abduction*: how, given the apparently infinite number of possible explanatory hypotheses in any given case, do scientists usually manage to hit upon the correct explanation so quickly? In some places in his writings, Peirce suggests that the apparent attunement between our abductive faculty, and the order of the universe may have an evolutionary explanation. More often than not though, he suggests that we simply have to hope that the universe is orderly, and that our intellectual faculties are attuned to that order, in order for the activity of rational inquiry to make any sense whatsoever. In other words, Peirce, like Plato, believes that the possibility of rational inquiry rests upon hopes in things that are not themselves amenable to rational demonstration.

The nature and legitimacy of hope in general, and the fundamental hopes that underlie the possibility of rational inquiry in particular, was the subject of chapter 2. We began with Aquinas' definition of hope, modified slightly, and expressed in the contemporary idiom of propositional attitudes[8]:

[5] See pp24-25.

[6] See pp11-19.

[7] See section 1.3.

[8] See section 2.2.

S hopes that p iff:

1. S desires that p *or* S believes that p is good.
2. S believes that p is uncertain.
3. S believes that p is possible *or* has no reason to believe that p is impossible.

However, further examination revealed that this definition was equally compatible with the presence of despair. This tells us something important about the nature of hope, namely, that as long as certain, somewhat minimal conditions are met, there is only a loose correlation between what someone believes to be the case and their adoption of an attitude of hope, rather than despair. I concluded that the adoption of one attitude rather than the other by different people in the same circumstances is therefore a reflection of individual temperament, as much as anything else.

This raised the general question of the relationship between hope and experience. Aquinas' definition supplied a few ideas[9]:

- Experience may suggest that a particular good is attainable; this permits and enables us to entertain hopes of attaining that good in the future.

- Experience may show that what we previously hoped for is impossible; the rational policy in these circumstances is to abandon all of our hopes in that direction.

- *Lack of* experience may cause us to persist in hoping for a particular good when it is in fact impossible.

But, we decided that these general conclusions are not much use when applied to the specific hopes that underwrite the possibility of rational inquiry, because:

1. They tell us little about when it might be legitimate to*act* on the basis of nothing more than a hope, as commitment to a life of inquiry demands.

[9] See pp46-50.

2. They deal with mundane hopes, where the relationship between hope and experience is relatively clear cut. The hopes that underlie the possibility of rational inquiry are not of this type.

For an account of the relationship between hope and action, we turned to the work of William James, and, in particular, to his essay, "The Will to Believe"[10]. We accepted James' conclusion that what he calls our "passional nature" may - indeed must - fill the breach left by reason where the issue concerned is live forced and momentous[11]. I argued that the decision to actively commit oneself to a life of rational inquiry meets James' criteria, and that we may therefore legitimately entertain the hopes that this commitment presupposes. In doing so, I argued that James' conclusions, although usually thought of as applying to beliefs, may also be applied to hopes, and that the kinds of cases that both James and I are interested in seem to occupy a grey area between belief and hope anyway.

Finally, in chapter 2, we turned to Kant for a more detailed account of the unusual character of the fundamental hopes that underlie the possibility of rational inquiry, and for further insight into when it might be legitimate to entertain them[12]. We saw that the hope that the universe is orderly and that our intellects are attuned to that order (or, more poetically, hope in the Theory of Recollection) has, as its object, something that lies beyond the bounds of possible experience, and that such hopes are only justified as long as they answer to the "needs of reason"[13]. Fortunately, the hopes that interest us do satisfy this requirement, and this provides further reassurance that they may legitimately be entertained.

Chapter 3 began work on the other main theme of the thesis, understanding, building upon the insights from the *Meno* dialogue described in chapter 1. I began by distinguishing between dynamic and static understanding: roughly speaking, the former is episodic, occurring at a specific time, whilst the latter is a state of the individual that persists over a period of time[14]. The two aspects of understanding are related, in that we may reasonably hope that a series of episodes of dynamic understanding will lead to static understanding, in the long run. The core of the chapter was then devoted to developing an account of dynamic

[10] See pp50-61.
[11] For an explanation of the meaning of these terms, see chapter 2, p52.
[12] See section 2.4.
[13] See chapter 2, p68ff, for an explanation of this term.
[14] For further discussion of this distinction, see pp80-83.

understanding. I concluded that this is exhibited whenever an individual responds to new information with non-minimal comprehension, as defined below[15]:

S has non-minimal comprehension of new information, P, with respect to her existing beliefs, *iff*

N1: $R_1 \neq \Delta_P (R_0)$

A1: S sees P as *warranting* the discrepancies between R_1 and R_0

A2: P *does* warrant the discrepancies between R_1 and R_0

Where: R_0 and R_1 are sets of responses given by S to members of Q, the set of all the questions she understands[16], before and after being informed of P; Δ_P is the function that replaces all and only S' responses to members of Q_P before being informed of P, with the *correct* answers to the members of Q_P, *on the assumption that P is true*; and Q_P is the set of all questions that can be answered correctly, given *only* the information P, and competence in the language.

The remainder of the chapter, and the thesis, was devoted to exploring the nature of *static* understanding[17]. I began by introducing the concept of a model[18]:

M is a model of O, relative to a set of triples $[Q_m, Q_o, T]$ *iff*:

1. In each triple, Q_m is a set of questions about M, and Q_o is a set of questions about O

2. T represents a set of procedures (algorithmic and/or heuristic), by means of which it is possible, though not necessarily easy, to translate any member of Q_o into a member of Q_m, and any answer to a member of Q_m into an answer to the corresponding member of Q_o.

3. T translates correct answers to members of Q_m into correct answers to members of Q_o.

[15] For the rationale for this definition, see section 3.4.

[16] Note that this refers to *linguistic* understanding, as opposed to non-linguistic; see p80, for further details.

[17] For the remainder of this conclusion, I shall drop the qualification "static".

[18] See section 3.5.

4. M enables us to answer at least one member of Q_o that *could* have been left unanswered during M's construction.

Of these conditions, the fourth is critical, as it distinguishes a genuine model from what I called an *oracle* or a *look-up table.*

We noted that a series of episodes of *minimal* comprehension[19] will only ever lead to the construction of a kind of internal look up table, rather than a model of the domain in question. This suggested that a series of episodes of *non*-minimal comprehension, as occur whenever dynamic understanding is exhibited, may enable one to construct a model of the object of one's inquiry. This led us, via the following argument, to entertain the hypothesis that one has understanding of a given domain to the extent that one has an adequate model of it:

1. The consistent exercise of non-minimal comprehension with respect to information about a given domain, enables one to construct an adequate model of that domain[20].

2. The consistent exercise of dynamic understanding in a given domain may reasonably be expected to lead to static understanding of that domain.

3. Dynamic understanding is exhibited whenever an episode of non-minimal comprehension occurs.

4. *Therefore*, one has static understanding of a given domain to the extent that one has an adequate model of it.

Chapters 4 and 5 aimed to provide further evidence for this conclusion.

Chapter 4, inspired by the prevalence of spatial metaphors in the language we use to describe our inquiries, presented the following argument from analogy[21]:

1. Spatial orientation and understanding have many features in common.

[19] Where S *minimally* comprehends a new piece of information, P, with respect to her existing beliefs, *iff* $R_1 = \Delta_P (R_0)$. See section 3.4 for further details.

[20] Or, at least, it is reasonable to *hope* that it will.

[21] See p134ff.

2. In the case of spatial orientation, we can account for the presence of these features by supposing that the organism has a map (usually a mental one) of its environment.

3. *Therefore*, it is likely that the presence of the same features in cases of understanding may be accounted for by the existence of a map-like representation of the domain in question.

In addition, I argued that a map is a model of the terrain it represents[22]; this being so, the conclusion that understanding a given domain is a matter of having an adequate model of it began to look very plausible.

In support of the first premise of the above argument, the first part of chapter 4 assembled a list of features characteristic of spatial orientation, that are also present when understanding appears to be exhibited. In each case[23]:

- The individual concerned has the ability to solve *many* problems.
- Success in solving problems is *non-accidental*.
- Solutions to problems are *efficient*.
- Solutions to problems are *flexible*.
- Solutions to problems are *robust*.
- Solutions to problems are *reversible*.

In looking for support for the second premise, I examined the difference between maps and route descriptions, concluding that an organism with a map like representation of its environment is, as one might expect, likely to exhibit the features characteristic of spatial orientation.

Chapter 5 approached the issue of the relationship between models and understanding from a different angle, aiming to show that current work in the philosophy of science supports the view that we understand a given domain to the extent that we have an adequate model of it.

[22] See p142ff.
[23] See section 4.3.

The first half of the chapter described the influential semantic[24] conception of scientific theories[25], and concluded that:

- At the level of *microstructure*[26], a scientific theory provides us with a variety of model types; these may be used to solve problems about real physical systems falling within the intended scope of the theory.

- At the level of *macrostructure*, a scientific theory is a *system* of model types, structured by genus-species and similarity relations; this system is itself a kind of large scale model of the theory's domain.

Since scientific theories are prime sources of scientific understanding, these conclusions support the claim that we understand a given domain to the extent that we have an adequate model of it.

In the second half of chapter 5, I turned to the nature of scientific explanation, considering three contemporary approaches to the issue: theoretical unification; causal-mechanical; and pragmatic[27]. I argued that:

- In each case, the idea that a scientific theory is a system of model types may be used to shed light on the nature of the approach concerned.

- The idea that a scientific theory is a system of model types helps us to reconcile the three approaches, since it enables us to see them as describing three complementary aspects of a single phenomenon.

Since it is usually acknowledged that there is a close connection between scientific explanation and scientific understanding, these conclusions provide further evidence for the claim that we understand a given domain to the extent that we have an adequate model of it.

[24] See chapter 5, p152ff for an explanation of this term, and its relationship to the syntactic conception of scientific theories.

[25] See section 5.2.

[26] See chapter 5, p160 for a discussion of the difference between the microstructure and the macrostructure of scientific theories.

[27] See section 5.4.

I stated in the introduction that this thesis would not attempt to provide a comprehensive philosophical theory of rational inquiry, such an undertaking being hopelessly ambitious for a work of this scale. I now wish to acknowledge just how modest the conclusions of the thesis are, for they provide only the starting point and the end point of such a comprehensive theory:

- It is the hope that the universe is orderly, and that our intellects are attuned to that order, that *starts* us on the path of inquiry, and sustains us if our resolve falters.

- The *end* of inquiry is understanding, which consists in having an adequate model of the object of the inquiry.

Of the long, mysterious road between the origin of inquiry in hope, and its terminus in understanding, I have said little in this piece of work. That must remain the subject of another inquiry.

Bibliography

Allen, R. E: "Anamnesis in Plato's *Meno* and *Phaedo*"; *Review of Metaphysics* 13, pp165-74.

Almog, *et. al.* (eds.): *Themes from Kaplan*; Oxford University Press, 1991.

Appleyard, D: "Structuring a City"; *Environment and Behaviour* 2, pp100-16.

Aquinas, St. Thomas: *Summa Theologiae*; Cambridge, Blackfriars, 1964 onwards.

Aristotle: *Metaphysics*; in J. Barnes (ed.), *The Complete Works of Aristotle: Revised Oxford Translation.*

Aristotle: *Posterior Analytics*, J. Barnes (tr.); Oxford University Press, 1993.

Ayim, M: *Peirce's View of the Role of Reason and Instinct in Scientific Inquiry*, Anu Prakashan, 1982.

Barnes, J. (ed.): *The Complete Works of Aristotle: Revised Oxford Translation*; Princeton University Press, 1984.

Barnes, J: "Socrates and the Jury: Paradoxes in Plato's Distinction Between Knowledge and True Belief (II)"; *Proceedings of the Aristotelian Society*, 1980, supplementary volume, pp193-206.

Beck, L. W (ed.): *Kant: On History*; New York, Bobbs-Merrill, 1963.

Benson, H: "Meno, the Slave Boy and the *Elenchos*"; *Phronesis* 35 (1990), pp128-58.

Berti, E. (ed.): *Aristotle on Science: The Posterior Analytics*; Podova, Editrice Antenore, 1981.

Bird, G. H: *William James: Selected Writings*; London, J. M. Dent, 1995.

Bird, G. H: *William James*; London, Routledge and Kegan Paul, 1986.

Board, C: "Maps as Models"; in R. J. Chorley and P. Haggett (eds.), *Models in Geography.*

Bridgman, P. W: *The Logic of Modern Physics*; New York, Macmillan, 1968.

Bromberger, S: "An Approach to Explanation"; in *On What We Know We Don't Know.*

Bromberger, S: "Rational Ignorance"; in *On What We Know We Don't Know.*

Bromberger, S: "Why-Questions"; in *On What We Know We Don't Know.*

Bromberger, S: "What We Don't Know When We Don't Know Why"; in *On What We Know We Don't Know.*

Bromberger, S: *On What We Know We Don't Know*; University of Chicago Press, 1992.

Brown, J. R: "Proofs and Pictures"; *British Journal of the Philosophy of Science* 48 (1997), pp161-80.

Bunge, W: *Theoretical Geography*; Lund, University of Lund Press, 1966.

Burnyeat, M: "Socrates and the Jury: Paradoxes in Plato's Distinction Between Knowledge and True Belief (I)"; *Proceedings of the Aristotelian Society*, 1980, supplementary volume, pp173-92.

Burnyeat, M: "Aristotle on Understanding Knowledge"; in E. Berti (ed.), *Aristotle on Science: The Posterior Analytics*.

Burnyeat, M: "Wittgenstein and Augustine *de Magistro*"; *Proceedings of the Aristotelian Society*, 1987, supplementary volume, pp1-24.

Butts, R. E: *Kant's Philosophy of Physical Science*; Dordrecht, D. Reidel, 1986.

Campbell, J: *Past, Space and Self*; Cambridge Mass., MIT Press, 1994.

Carnap, R: *Logical Foundations of Probability* (Second Edition); Chicago, University of Chicago Press, 1962.

Cavallar, G: "For What May I Hope?"; *Kant Studien* 85 (1994), pp356-58.

Chomsky, N: "On the Nature, Use and Acquisition of Language"; in W. Lycan (ed.), *Mind and Cognition: A Reader*.

Chorley, R. J. and Haggett, P. (eds.): *Models in Geography*; London, Methuen and Co., 1967.

Clifford, W. K: *The Ethics of Belief and Other Essays*, L. Stephen and F. Pollock (eds.); London, Watts and Co., 1947.

Clifford, W. K: "The Ethics of Belief"; in *The Ethics of Belief and Other Essays*.

Coffa, A: "Hempel's Ambiguity"; *Synthese* 28 (1974), pp141-63.

Cooper, N:"The Epistemology of Understanding"; *Inquiry* 38 (1995), pp205-15.

Cooper, N: "Understanding"; *Proceedings of the Aristotelian Society* 1994, supplementary volume, pp1-26.

Crease, R. P. and Mann, C. C: *The Second Creation*; London, Quartet Books, 1997.

Crombie. I. M: "Socratic Definition"; in J. M. Day (ed.), *Plato's Meno in Focus*.

Day, J. M: *Plato's Meno in Focus*; London, Routledge, 1994.

Day, J. P: "Anatomy of Hope and Fear"; *Mind* 79 (1970), pp369-84.

Day, J. P: "Hope"; *American Philosophical Quarterly* 6 (1969), pp89-102.

Dennett, D: *Darwin's Dangerous Idea*; London, Penguin Books, 1995.

Desjardins, R: "Plato's *Meno*"; *review of Metaphysics* 39, pp261-81.

Devereux, D: "Nature and Teaching in Plato's *Meno*"; *Philosophy* 23, pp118-26.

Dewey, J: *Logic: The Theory of Inquiry*; London, George Allen and Unwin, 1939.

Downie, R. S: "Hope"; *Philosophy and Phenomenological Research* 24 (1963-64), pp248-50.

Downs, R. M. and Stea, D: *Image and Environment: Cognitive Mapping and Spatial Behaviour*; Chicago, Aldine, 1973.

Downs, R. M. and Stea, D: *Maps in Mind: Reflections on Cognitive Mapping*; London, Harper and Row, 1977.

Downs, R. M: "Maps and Metaphors"; *The Professional Geographer* 33 (1981), pp287-93.

Dray, W: *Laws and Explanation in History*; New York, Oxford, 1964.

Dretske, F: *Knowledge and the Flow of Information*; Oxford, Basil Blackwell, 1981.

Evans, G: *The Varieties of Reference*; Oxford University Press, 1982.

Evans, G: "Understanding Demonstratives"; in P. Yourgrau (ed.), *Demonstratives*.

Feigl, H. and Sellars, W. (eds.): *Readings in Philosophical Analysis*; New York, Appleton-Century-Crofts, 1949.

Feigl, H: "Some Remarks on the Meaning of Scientific Explanation"; in H. Feigl and W. Sellars (eds.), *Readings in Philosophical Analysis*.

Fine, G: "Inquiry in the *Meno*"; in R. Kraut (ed.), *The Cambridge Companion to Plato*.

Franklin, R. L: "Knowledge, Belief and Understanding"; *The Philosophical Quarterly* 31 (1981), pp193-208.

Friedman, M: "Explanation and Scientific Understanding"; *Journal of Philosophy* 71 (1974), pp5-19.

Friedman, M: "Theoretical Explanation"; in R. Healey (ed.), *Reduction, Time and Reality*.

Garfinkel, A: *Forms of Explanation*; New Haven, Yale University Press, 1981.

Geach, P. T: "Plato's *Euthyphro*: An Analysis and Commentary"; *Monist* 50 (1966), pp369-82.

Gettier, E. L: "Is Justified True Belief Knowledge?"; *Analysis* 23 (1963), pp121-23.

Godfrey, J. J: *A Philosophy of Human Hope*; Dordrecht, Nijhoff, 1987.

Gooch, P. W: "Irony and Insight in the *Meno*"; *Laval* 43, pp189-204.

Grice, P: *Studies in the Way of Words*; Cambridge Mass., Harvard University Press, 1989.

Grice, P: "Logic and Conversation"; in *Studies in the Way of Words*.

Grimaldi, N: "Hope and Despair of Reason in Kant"; *Kant Studien* 82 (1991), pp129-45.

Grimaltos, T. and Hookway, C: "When Deduction Leads to Belief"; *Ratio* 8 (1995), pp24-41.

Guyer, P: *The Cambridge Companion to Kant*; Cambridge University Press, 1996.

Hanson, N. R: *Patterns of Discovery*; Cambridge University Press, 1958.

Healey, R. (ed.): *Reduction, Time and Reality*; Cambridge University Press, 1981.

Hempel, C. G., and Oppenheim, P: "Studies in the Logic of Explanation"; *Philosophy of Science* 15 (1948), pp135-75.

Hempel, C. G: "Aspects of Scientific Explanation"; in*Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.*

Hempel, C. G: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*; New York, Free Press, 1965.

Hesse, M: *Models and Analogies in Science*; Notre Dame, University of Notre Dame Press, 1966.

Hoerber, R. G: "Plato's *Meno*"; *Philosophy* 5, pp78-102.

Hookway, C: *Peirce*; London, Routledge, 1985.

Hookway, C: *Scepticism*; London, Routledge, 1990.

Hookway, C: *Truth, Rationality and Pragmatism: Themes from Peirce*; Oxford University Press, 2000.

Hookway, C: "Metaphysics, Science and Self-Control"; in *Truth, Rationality and Pragmatism: Themes from Peirce.*

Houser, N. and Kloesel, C. (eds.): *The Essential Peirce, Volume 1 (1867-1893)*; Bloomington and Indianapolis, Indiana University Press, 1992.

Hume, D: *An Enquiry Concerning Human Understanding*; Indianapolis, Hackett Publishing Company, 1993.

Huxley, T. H: *Darwiniana*; New York, AMS Press, 1970.

Irwin, T: "Recollection and Plato's Moral Theory"; *Review of Metaphysics* 27, pp752-72.

Irwin, T. and Fine, G: *Aristotle: Selections*; Indianapolis, Hackett, 1995.

Ittelson, W. H: "Environment Perception and Contemporary Perceptual Theory"; *Transactions of the New York Academy of Sciences* 32 (1970), pp807-15.

James, W: "The Will to Believe"; in G. H. Bird (ed.), *William James: Selected Writings.*

James, W: "The Sentiment of Rationality"; in G. H. Bird (ed.), *William James: Selected Writings.*

James, W: "The Dilemma of Determinism"; in G. H. Bird (ed.), *William James: Selected Writings.*

Jeffrey, R: "Statistical Explanation vs Statistical Inference"; in N. Rescher (ed.), *Essays in Honour of Carl G. Hempel.*

Jones, T: "Reductionism and the Unification Theory of Explanation"; *Philosophy of Science* 62 (1995), pp21-30.

Kant, I: *Cambridge Edition of the Works of Immanuel Kant*; Cambridge University Press, 1992 onwards.

Kant, I: *Critique of Judgement*, W. S. Pluhar (tr.); Indianapolis, Hackett, 1987.

Kant, I: *Critique of Practical Reason*, Lewis White Beck (tr.); Indianapolis, Bobbs-Merrill, 1976.

Kant, I: *Critique of Pure Reason*, N. Kemp-Smith (tr.); London, Macmillan, 1929.

Kant, I: *Dreams of a Spirit Seer*; in *Cambridge Edition of the Works of Immanuel Kant*, volume 1.

Kant, I: *Groundwork of the Metaphysic of Morals*, J. W. Ellington (tr.); Indianapolis, Hackett, 1981.

Kant, I: "The End of All Things"; in L. W. Beck (ed.), *Kant: On History*.

Kant, I: "What is Orientation in Thinking?"; in H. Reiss (ed.), *Kant: Political Writings*.

Kaplan, D: "Demonstratives"; in Almog, *et. al.*(eds.), *Themes from Kaplan*.

Kitcher, P. and Salmon, W. C. (eds.): *Scientific Explanation*; Minneapolis, University of Minnesota Press, 1989.

Kitcher, P: "Explanation, Conjunction and Unification"; *Journal of Philosophy* 73, pp207-12.

Kitcher, P: "Explanatory Unification and the Causal Structure of the World"; in P. Kitcher and W. C. Salmon (eds.), *Scientific Explanation*.

Kitcher, P: "Explanatory Unification"; *Philosophy of Science* 48 (1981), pp507-31.

Kitcher, P: "Projecting the Order of Nature"; in R. E. Butts (ed.), *Kant's Philosophy of Physical Science*.

Kitcher, P: "Two Approaches to Explanation"; *Journal of Philosophy* 1985, pp632-39.

Klein, J: *A Commentary on Plato's Meno*; University of Chicago Press, 1989.

Kraut, R. (ed.): *The Cambridge Companion to Plato*; Cambridge University Press, 1992.

Kuhn, T: *The Structure of Scientific Revolutions*; University of Chicago Press, 1962.

Kuipers, B: "The 'Map in the Head' Metaphor"; *Environment and Behaviour* 14 (1982), pp202-20.

Lakatos, I: *Proofs and Refutations*, J. Worrall and E. Zahar (eds.); Cambridge University Press, 1976.

Lawrence, N. and O'Connor, D. (eds.): *Readings in Existential Phenomenology*; Prentice-Hall, 1967.

Lear, J: *Aristotle: The Desire to Understand*; Cambridge University Press, 1988.

Lipton, P: *Inference to the Best Explanation*; London, Routledge, 1991.

Lloyd, G. E. R: "The *Meno* and the Mysteries of mathematics"; *Philosophy*, 1992, pp166-83.

Locke, J: *An Essay Concerning Human Understanding*; P. Nidditch (ed.); Oxford, Clarendon Press, 1975.

Lycan, W. (ed.): *Mind and Cognition: A Reader*; Oxford, Basil Blackwell, 1990.

Lynch, K: *The Image of the City*; Cambridge, MIT Press, 1960.

McDermott, T: *St. Thomas Aquinas' Summa Theologiae: A Concise Translation*; London, Methuen, 1989.

Marshall, J. D: "On What We May Hope"; *Studies in Philosophy and Education* 13, pp307-24.

Marcel, G: *Homo Viator: Introduction to a Metaphysics of Hope*; London, Victor Gollancz, 1951.

Marcel, G: "Desire and Hope"; in N. Lawrence and D. O'Connor (eds.), *Readings in Existential Phenomenology*.

Mates, B: *The Sceptic Way*; Oxford University Press, 1996.

Menninger, K: "Hope"; *American Journal of Psychiatry* 116 (1959), pp481-91.

Mill, J. S: *Nature, the Utility of Religion, and Theism*; London, Watts & Co., 1904.

Mill, J. S: *Theism*; in *Nature, the Utility of Religion, and Theism.*

Moline, J: "Meno's Paradox?"; *Philosophy* 14, pp153-61.

Moltmann, J: *Theology of Hope*; London, SCM Press, 1967.

Moore, A. W: *The Infinite*; London, Routledge, 1981.

Moore, A. W: "Ineffability and Reflection: An Outline of the Concept of Knowledge"; *European Journal of Philosophy* 1 (1993), pp285-308.

Moravcsik, J: "Understanding"; *Dialectica* 33 (1979), pp201-15.

Morgenbesser, S (ed.): *Philosophy of Science Today*; New York, Basic Books, 1967.

Morris, C: "Foundations of the Theory of Signs"; in O. Neurath, R. Carnap, and C. Morris (eds.), *Foundations of the Unity of Science: Towards an International Encyclopaedia of Unified Science.*

Muyskens, J. L: "Religious Belief as Hope"; *International Journal of the Philosophy of Religion* 5, pp246-53.

Nagel, E: *The Structure of Science*; New York, Harcourt, 1961.

Nehamas, A: "Meno's Paradox and Socrates as a Teacher"; in J. M. Day (ed.), *Plato's Meno in Focus*.

Neurath, O., Carnap, R. and Morris, C. (eds.): *Foundations of the Unity of Science: Towards an International Encyclopaedia of Unified Science*; Chicago University Press, 1955.

O' Keefe, J. and Nadel, L: *The Hippocampus as a Cognitive Map*; Oxford, Clarendon Press, 1978.

Parrett, H., and Bouveresse, J. (eds.): *Meaning and Understanding*; Berlin/New York, de Gruyter, 1981.

Peirce, C. S: *The Collected Papers*, C. Hartshorne and P. Weiss (eds.); Cambridge, Harvard University Press, 1935.

Peirce, C. S: "The Fixation of Belief"; in N. Houser and C. Kloesel (eds.), *The Essential Peirce, Volume 1 (1867-1893*.

Peirce, C. S: "How to Make Our Ideas Clear"; in N. Houser and C. Kloesel (eds.), *The Essential Peirce, Volume 1 (1867-1893*.

Perry, J: "Frege on Demonstratives"; in P. Yourgrau (ed.), *Demonstratives*.

Perry, J: "The Problem of the Essential Indexical"; in J. Perry, *The Problem of the Essential Indexical*

Perry, J: *The Problem of the Essential Indexical*; Oxford University Press, 1993

Plato: *Protagoras and Meno*, W. C. K. Guthrie (tr.); London, Penguin Classics, 1956.

Powers, L: "Knowledge by Deduction"; *The Philosophical Review* 87 (1978), pp337-71.

Radford, C: "Hoping and Wishing"; *Proceedings of the Aristotelian Society*, supplementary volume XLIV (1970).

Railton, P: "A Deductive-Nomological Model of Probabilistic Explanation"; *Philosophy of Science* 45, pp206-26.

Railton, P: "Explanation and Metaphysical Controversy"; in W. C. Salmon, *et. al.* (eds.), *Scientific Explanation*.

Railton, P: "Probability, Explanation and Information"; *Synthese* 48 (1981), pp233-56.

Redhead, M: "Models in Physics"; *British Journal of the Philosophy of Science* 31 (1980) pp145-63.

Reiss, H (ed.): *Kant: Political Writings*, H. B. Nisbet (tr.); Cambridge University Press, 1990.

Rescher, N. (ed.): *Essays in Honour of Carl G. Hempel*; Dordrecht, D. Reidel, 1969.

Rescher, N: *Peirce's Philosophy of Science*; Notre Dame University Press, 1978.

Resnik, M. D. and Kushner, D: "Explanation, Independence and Realism in Mathematics"; *British Journal for the Philosophy of Science* 38 (1987), pp141-58.

Richards, P: "Kant's Geography and Mental Maps"; *Transactions of the Institute of British Geographers* 61, pp1-16.

Rosenberg, J. F: "On Understanding the Difficulty in Understanding Understanding"; in H. Parrett and J. Bouveresse (eds.), *Meaning and Understanding*.

Ruben, D: *Explanation*; Oxford University Press, 1993.

Russell, B: "Pragmatism"; *Edinburgh Review*, 1909.

Salmon, W. C: *Four Decades of Scientific Explanation*; Minneapolis, University of Minnesota Press, 1989.

Salmon, W. C: *Scientific Explanation and the Causal Structure of the World*; Princeton University Press, 1984.

Salmon, W. C. *et. al.* (eds.): *Scientific Explanation* (Minnesota Studies in the Philosophy of Science, volume XIII); Minneapolis, University of Minnesota Press, 1989.

Salmon, W. C: "Conflicting Conceptions of Scientific Explanation"; *Journal of Philosophy* 1985, pp651-54.

Salmon, W. C. and Kitcher, P: "Van Fraassen on Explanation"; in D. Ruben (ed.), *Explanation*.

Schuler, J. A: "Reasonable Hope: Kant as Critical Theorist"; *History of European Ideas* 21 (1995), pp527-33.

Scott, D: *Recollection and Experience: Plato's Theory of Learning and its Successors*; Cambridge University Press, 1995.

Scott, D: "Platonic Anamnesis Revisited"; *Classical Quarterly* 37 (1987), pp346-66.

Sextus Empiricus: *Outlines of Pyrrhonism*, S. G. Etheridge (tr.); Indianapolis, Hackett Publishing Co., 1985.

Shannon, B: "*Meno* - A Cognitive Psychological View"; *British Journal of the Philosophy of Science* 35 (1984), pp129-47.

Sintonen, M: "Explanation: In Search of the Rationale"; in W. C. Salmon, *et. al.* (eds.), *Scientific Explanation*.

Skorupski, J: *Mill*; London, Routledge, 1991.

Steiner, M: "Mathematical Explanation"; *Philosophical Studies* 34 (1978), pp135-51.

Stratton-Lake, P: "Reason, Appropriateness and Hope: Sketch of a Kantian Account of Finite Rationality"; *International Journal of Philosophical Studies* 1, pp61-80.

Suckiel, E. K: *The Pragmatic Philosophy of William James*; Notre Dame University Press, 1982.

Suppe, F (ed.): *The Structure of Scientific Theories*; University of Illinois Press, 1977.

Suppes, P: "What is a Scientific Theory?"; in S. Morgenbesser (ed.), *Philosophy of Science Today*.

Sutherland, S: "Hope"; *Philosophy* 25 (1989), pp193-206.

Tigner, S. S: "On the Kinship of all Nature"; *Philosophy* 15, pp1-4.

Tiles, J: *Dewey*; London, Routledge and Kegan Paul, 1988.

Toulmin, S: *Foresight and Understanding*; London, Hutchinson, 1961.

Toulmin, S: *Philosophy of Science*; London, Hutchinson & Co., 1967.

Unger, P: "On Experience and the Development of the Understanding"; *American Philosophical Quarterly* 3 (1966), pp48-56.

van Fraassen, B. C: *Laws and Symmetries*; Oxford University Press, 1989.

van Fraassen, B. C: *The Scientific Image*; Oxford University Press, 1980.

van Fraassen, B. C: "On the Extension of Beth's Semantics of Physical Theories"; *Philosophy of Science* 1970, pp325-39.

van Fraassen, B. C: "Salmon on Explanation"; *Journal of Philosophy* 1985, pp639-51.

Vlastos, G: "What does Socrates Understand by his 'What is F?' Question"; in *Platonic Studies*.

Vlastos, G: *Platonic Studies*; Princeton University Press, 1981.

Wartenberg, T. A: "Reason and the Practice of Science"; in P. Guyer (ed.), *The Cambridge Companion to Kant*.

Wheatley, J. M. O: "Wishing and Hoping"; *Analysis* 18 (1958).

Williams, B: *Problems of the Self*; Cambridge University Press, 1973.

Williams, B: "Deciding to Believe"; in *Problems of the Self*.

Wittgenstein, L: *Philosophical Investigations*; Oxford, Basil Blackwell, 1958.

Wood, D: "Rational Theology"; in P. Guyer (ed.), *The Cambridge Companion to Kant*.

Woodward, J: "The Causal Mechanical Model of Explanation"; in W. C. Salmon, *et. al.* (eds.), *Scientific Explanation*.

Yourgrau, P (ed.): *Demonstratives*; Oxford University Press, 1990.

Zagzebski, L: *Virtues of the Mind*; Cambridge University Press, 1996.