PRAGMATICS & RATIONALITY

Nicholas Elwyn Allott

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University College London, September 2007

The work in this dissertation is entirely my own.

Nicholas Allott

ABSTRACT

This thesis is about the reconciliation of realistic views of rationality with inferential-intentional theories of communication.

Grice (1957; 1975) argued that working out what a speaker meant by an utterance is a matter of inferring the speaker's intentions on the presumption that she is acting rationally. This is abductive inference: inference to the best explanation for the utterance. Thus an utterance both rationalises and causes the interpretation the hearer constructs.

Human rationality is bounded because of our 'finitary predicament': we have limited time and resources for computation (Simon, 1957b; Cherniak, 1981). This raises questions about the explanatory status of inferential-intentional pragmatic theories. Gricean derivations of speakers' intentions seem costly, and generally hearers are not aware of performing explicit reasoning. Utterance interpretation is typically fast and automatic. Is utterance interpretation a species of reasoning, or does the hearer merely act *as if* reasoning?

Within the framework of cognitive science, mental processing is understood as transitions between mental representations. I develop a traditional view of rationality as reasoning ability, where this is essentially the ability to make transitions that preserve rational acceptability. Following Grice (2001), I claim that there is a 'hard way' and a 'quick way' of reasoning. Work on bounded rationality suggests that much cognitive work is done by heuristics, processes that exploit environmental structure to solve problems at much lower cost than fully explicit calculations. I look at the properties of heuristics that find solutions to open-ended problems such as abductive inference, particularly sequential search heuristics with aspiration-level stopping rules.

I draw on relevance theory's view that the comprehension procedure is a heuristic which exploits environmental regularities due to utterances being offers of information (Sperber & Wilson, 1986). This kind of heuristic, I argue, is the 'quick way' that reasoning proceeds in utterance interpretation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am profoundly grateful to my supervisor, Deirdre Wilson for her unfailing support, understanding, encouragement and patience. She has been the greatest influence on the views I express here, through our discussions, her writings and lectures and her work with Dan Sperber. No one could wish for a better supervisor with a better grasp of what matters.

I would not have been in linguistics or in the UCL linguistics department without Neil Smith. I thank him for taking me onto the MA course (nothing has been the same since), for formative discussions, and for considerable support and encouragement.

I am privileged, and very grateful, to have received funding for my PhD studies, first from the Department of Phonetics and Linguistics, UCL (2001–2003) and then from the Arts and Humanities Research Council (2003-2006).

During the MA and some years of my PhD I was lucky enough to be able to attend lectures and classes by Misi Brody, Robyn Carston, John Harris, Richard Horsey, Corinne Iten, Geoff Lindsey, Ad Neeleman, Kriszta Szendrői, Hans van de Koot and Tim Wharton as well as Neil and Deirdre. I thank them all for a perfect start in linguistics.

I have also been lucky to find myself among relevance theorists and UCL linguists, two overlapping groups of highly intelligent and thoughtful people. For illuminating discussions I am grateful to (among others, and I apologise to those I have missed out) Richard Breheny, Coralie Chevallier, Billy Clark, Annabel Cormack, Ingrid Lossius Falkum, Alison Hall, Vikki Janke, Mark Jary, Napoleon Katsos, Patricia Kolaiti, Eleni Kriempardis, Anna Pollard, George Powell, Paula Rubio Fernández, Louis de Saussure, Hitoshi Shiraki, Hanna Stöver, Marco Tamburelli, Rob Truswell, Rosa Elena Vega Moreno, Reiko Vermeulen and Vlad Žegarac.

I particularly want to thank Dirk Bury for discussions and for his wonderful classes on Misi's syntax, Tim Wharton for passing on enthusiasm and knowledge about Grice, and Milena Nuti for discussions about all sorts of things. I owe a huge debt of gratitude to Hiroyuki Uchida for his help with a great deal of my work, teaching as well as research, and for numerous discussions.

Thanks to Marc Richards and to Mark Textor for their friendship as well as their thoughts. I have neglected good friends writing this thesis. I particularly thank Nick Doody for his support and being a good friend over many years.

My mother and father have always put their children first. I am very proud of them and of Kate and Lucy and I thank them all for their love and support.

I would be more grateful to Teddy, Smokey, Spot, Clara and Alice for inspiring some of the examples in this thesis if I had not spent half this summer rescuing frogs they had captured.

Finally, I thank my wife, Jui Chu, for her understanding, bravery and love.

All the mistakes in this thesis are my own, of course.

Chapter 1 · Introduction

1.1 PRAGMATICS, RATIONALITY AND COGNITION

Why study pragmatics and rationality together? If I address an utterance to you, you have understood if you grasp what I meant by the utterance. Part of this is a matter of recognising the linguistic items used and any non-linguistic gestures which encode meanings[1]. If you know the meanings of the linguistic and non-linguistic components of an utterance, then you have a better chance of working out what I meant. As Grice stressed, however, what a speaker means by an utterance typically goes well beyond what the phrase uttered[2] encodes[3]. The meaning conveyed by uttering a phrase varies depending on how it is uttered and in what context. Understanding utterances involves inference that takes into account these factors and the linguistic meaning of the phrases uttered. Pragmatics is the study of this inferential aspect of utterance understanding and production.

Any understanding of what a speaker means by the sounds she makes, the way that she waves her hands around, and so on, relies on two assumptions: 1) that she is behaving rationally, so that her behaviour serves her purposes, or is at

[1] Non-linguistic gestures divide into those which encode meaning, such as thumbs up for 'OK', and those which do not have any encoded meaning and are invented and understood purely inferentially. See Sperber & Wilson, 1986, p. 52 for examples. A further division can be made between gestures (and sounds) with natural and those with non-natural meaning. See Wharton, 2003 for comprehensive discussion.

[2] I write 'phrase uttered' rather than 'sentence uttered' because many utterances are of less than complete sentences (Barton, 1990; Progovac, Paesani, Casielles, & Barton, 2006). There is reason to believe that most are complete linguistic constituents, that is, phrases.

[3] Grice's distinction was "between what is said (in a favoured sense) and what is implicated" (Grice, 1989c, p. 41). I follow relevance theory and other 'radical pragmatics' in believing that inference (often, indeed typically) contributes to the explicit meaning of an utterance including the proposition expressed (Wilson & Sperber, 1981; Carston, 1988). Hence the neutral formulation in the text at this point.

least intended to, and 2) that she intends to convey meaning. If she does not intend to communicate she might be making noises and waving her hands around for some other reason – to scare away a fly, or just to amuse herself. If she is not behaving rationally at all then it will be hard to infer anything about her intentions. So communication can only get going if the hearer can assume the speaker is rational, and is rational himself to the degree that he is capable of working out how the use of phrases and gestures is intended to serve the speaker's purposes. Looking at it from the other side, the speaker's production of phrases and gestures in order to get her meaning across to the hearer makes no sense unless there is a standing assumption on her part that the hearer is at least rational enough to be able to grasp that she meant something by her utterance and to have a good chance of working it out.

I am using the term 'pragmatics' to mean the study of this aspect of communication[4] on the assumption that this is a distinct task for the mind/brain from linguistic processing.

1.1.1 PROCESSING

There are strong reasons to think that pragmatic processing should be distinguished from linguistic processing. Linguistic items encode meaning, so linguistic processing is a matter of coding and decoding. What a speaker expresses by an utterance goes beyond what the utterance encodes, so working out the non-encoded information is a qualitatively different task from parsing. Indeed, the system can also work with gestures or sounds that do not encode any meaning. Thus pragmatic processing and parsing are conceptually quite separate (Sperber & Wilson, 1986). (See §1.3.1 below).

The strongest evidence that abilities are underpinned by different mental equipment is double dissociation, cases of selective impairment of each ability (Smith, 1999, p. 21). There is strong evidence that linguistic skills and pragmatic ability doubly dissociate. There are people who can manage language but not pragmatic processing. Others may have good pragmatic skills, but severely impaired language. (See §2.4.2 below.)

[4] In saying this I am agreeing with Sperber and Wilson (1986) on the issue of the semantics/ pragmatics borderline.

My thesis is that pragmatic processing is a bounded reasoning process, inferential and heuristic, which works so that an utterance made is both a cause of and a reason for the construction of a particular interpretation in the hearer's mind. I explain this thesis and the terms used in what follows: schematically first, in this introduction, then in greater detail.

1.1.2 GRICE'S THEORY

In the picture Grice developed (Grice, 1989c, Chapters 1-7, 14 and 18 and 'Retrospective Epilogue'), understanding what a speaker means by an utterance is a matter of recognising intentions that she has expressed. This insight provides the basis of an 'inferential model of communication' (the phrase and the observation are from Sperber & Wilson, 2004, p. 607). In this model, hearers understand utterances by inferring non-demonstratively what the speaker intended to convey on the basis of the linguistic meaning of the phrase uttered (if any) and other clues in the utterance and the context.

As Levinson says:

> Grice's theory gives us an account both of how we can communicate without conventional signals at all[5]... and of how we can communicate something distinct from what the conventional signals actually mean. (Levinson, 2006, p. 50)

How does this work? What makes the clues provided by a speaker reliable guides to her intended meaning? Grice proposed that conversation should be understood as cooperative, rational behaviour, suggesting that principles guiding conversation (the 'conversational maxims') should be derivable from the assumption that conversation is a cooperative activity carried out rationally (Grice, 1975), and that the standing assumption that those principles will be followed by speakers underwrites hearer's inferences about speaker meaning. Thus Gricean explanations of the derivation of inferred components of utterance meaning involve inference schemas similar to logical arguments: e.g. the speaker has said *x*, but *x* on its own does not meet the standards in

[5] Levinson has in mind here cases where the utterance is of a gesture or sound with no encoded meaning.

force in communication (perhaps it is not informative enough, or not relevant enough); the best explanation for that is that she meant *y*; she knew that I knew (etc.) that that was the best explanation, so she has intentionally communicated *y*.

So the idea that conveying meaning is a rational activity is the keystone of Gricean pragmatics, but one that Grice did not quite fix in place, for two reasons. The first is that he was not able to derive the Cooperative Principle and conversational maxims from considerations of rationality. My opinion is that rationality rather than cooperation is the key to pragmatics, agreeing in this respect, although not others, with Kasher (1976) and Horn (2006, p. 35) and *pace* Levinson who stresses cooperation over rationality (e.g. Levinson, 2006) (although he means something broader by cooperation than Grice did). I also think that the maxims are not derivable from any set of plausible assumptions.[6]

There is a second way in which Grice did not complete the linkage of communication and rationality, and this is the focus of my thesis. Even supposing that there are regulative principles for communication (whether Gricean maxims or otherwise) which will make inferential derivations of speaker meaning go through, there is still an important question about this kind of account. Do hearers really engage in this kind of reasoning? We are not generally aware of doing so, and there is a good deal of evidence that in reasoning we often use shortcuts. How are such schemas explanatory, then? How does it help us to understand communication to describe it in a way that is not a description of the mental states that must be gone through in order to understand an utterance?

It might seem that Gricean explanations of this type assume a classical, idealised vision of human rationality, since they explain behaviour in terms of an argument that one might construct, given ample leisure, to justify a judgment or decision that is actually made quickly. In fact, a Gricean picture of communication, I will argue, is compatible with a realistic view of human rationality as bounded by our limited mental resources.

[6] Kasher (1976; 1982) attempts a derivation of the conversational maxims from a rationality principle plus some other assumptions.

1.1.3 COOPERATION OR COORDINATION?

As I have said, what is meant can be inferred on the assumption that the speaker is acting rationally in making her utterance, since this allows the hearer to assume that the speaker has attempted to make effective use of the means chosen, that is, making an utterance at all. However, the hearer is not justified in assuming that the speaker has made or attempted to make the very best use of the means available. Human beings have only limited computational resources, and communication takes place quickly. Further, the speaker proceeds on the assumption that the hearer is able to make inferences – more specifically that he has the ability to infer her meaning from what is uttered and how it is uttered. This can only work if the utterance is suitable for the hearer to work out quickly, with finite resources.

There is a parallel with a game of catch. If I throw a ball to you, wanting you to catch it, and I am rational, I will try to make the trajectory suitable: towards the place where you will be, not too fast, nor too high. All of this can only happen if I assume, tacitly, that you have certain abilities.

Similarly, the receiver of a ball is justified in making certain assumptions about a ball that is apparently thrown with the intention that he catch it. The thrower should not intend to throw the ball too high or too fast for someone of the receiver's ability to catch – although she might by mistake, or if she does not really want the ball to be caught.

There are a number of similarities with a communicative situation. If you want me to understand you (and you are rational), you have to try to produce an utterance that I will be able to understand, just as if you want me to catch a ball you have to try to throw it so I can catch it. If I think that you want me to understand you, then I am rationally justified in assuming that you will try to produce an utterance that can be understood without excessive effort, just as, if I think that you want me to catch the ball you are throwing then I can assume you will try to produce a catchable throw, but cannot assume you will manage perfection, or even attempt it. I can assume that your throw will not require me to fling myself full-length at the ball like a slip-fielder, but I cannot assume that it will land in my hand with no effort on my part.

Does understanding utterances require cooperation? It need not. We are not as ready to catch balls as we are to understand utterances. If we were, I could throw a ball to any passer-by with the reasonable expectation of a catch, with no need to signal to him what kind of interaction we are engaging in except by the act of throwing, just as I can address an utterance to him and be understood on the basis of my attempt to be understood and his to understand. There does not need to be any shared aim in either case. I intend him to catch the ball or to understand the utterance and he may try to do so[7]. The speaker and hearer are engaged in different activities, with different aims, and this is not cooperation in the Gricean sense, which requires that the talk-exchange have a shared aim or purpose.

It is helpful to distinguish between coordination and cooperation. In cooperation two or more agents have a shared aim or purpose. Coordination is behaviour of two or more agents which dovetails so that it might appear that there is a joint purpose, whether there is or not. It has been one of the more useful roles of game theory to draw attention to coordinated behaviour that is not cooperative but emerges from quite separate, even conflicting aims pursued by interacting agents.

As far as communication is concerned, wanting to be understood does not entail wanting to contribute to a joint undertaking. Levinson separates these two aspects (although for him the distinction is between two types of cooperation):

> Interaction is by and large cooperative ... there is some level, not necessarily at the level of ulterior motivation, at which interactants intend their actions (a) to be interpretable (the underlying intentions to be recoverable), (b) to contribute to some larger joint undertaking (having a conversation, making a hut, even having a quarrel!) (2006, p. 4).

While I agree that speakers generally intend their utterances to be interpretable, I regard that as coordination arising from general rationality considerations, not cooperation. It is rarely worth saying something to someone else unless you intend to be understood. (There *are* cases where the speaker does not want to be understood or does not care whether she is: for example,

[7] It is not easy – perhaps impossible – to stop oneself from interpreting an utterance simply by choosing not to try. One *can* stop oneself from interpreting utterances by directing one's attention elsewhere preemptively, as when one reads a book to avoid eavesdropping on a conversation.

showing off by speaking in a foreign language not known to the hearer; but that is not communication.) So making an utterance with the aim or intention of being understood is not cooperation, but simply the rational interest in one's action succeeding (as argued in Sperber & Wilson, 1986, pp. 161–162; Sperber & Wilson, 1995, pp. 267–268). For a communicative action to succeed as such, it must interpretable.

Levinson's second sense in which interaction is cooperative is the sense Grice intended: that conversations or talk-exchanges must have an "a common purpose or set of purposes, or at least a mutually accepted direction" (Grice, 1975, p. 45). If this were true then communicative interaction would be genuinely cooperative. There are counterexamples which are nonetheless central cases of communication. I do not have the space here to go into this debate in detail but I give a few examples (see also Grice's own later discussion of the issue: Grice, 1989b, pp. 368–370). In interrogation and cross-examination, the participants' purposes may be diametrically opposed. The lawyer wants the defendant to incriminate himself or to appear unreliable or untruthful. The defendant wants the exact opposite. A second kind of case is one-off communication. A speaker making a one-off statement to a passer-by need share no purpose with the passer-by. Some threats and orders ('Get off the grass, or I'll belt you!') are non-cooperative in both ways: there is no established talk-exchange, so no pre-established purpose; and the speaker's purpose in making the utterance is to get the hearer to behave in a way that he would rather not, and has little in common with any purpose the hearer is likely to have. The hearer need not have any goal beyond the usual one of understanding what has been said to him (and this not explicitly or consciously – rather, it is built in to our pragmatics faculty, in my opinion). I take it, then, that communication is coordinative but not necessarily cooperative.

1.1.4 REALISTIC AND UNREALISTIC VIEWS OF RATIONALITY

According to Grice, utterances are actions directed towards fulfilling certain intentions. Rationality demands that action be appropriate to the desired end. Appropriateness implies efficiency: an action which will achieve the desired end but at much greater cost than an alternative is not as appropriate as that alternative, other things being equal. Speakers will not make utterances perfect at all costs, but must put in enough effort to make their utterance effective. Utterances also demand effort of hearers, since they must work out what was meant.

Considerations of this sort have led to pragmatic theories that are broadly Gricean but suggest that a balance is struck between the effort required by an utterance and the effects produced by it. Kasher advocates a *principle of effective means*: "Given a desired end, one is to choose that action which most effectively, and at least cost, attains that end, ceteris paribus." (1982) Horn collapses Grice's maxims into two principles: one that requires the production of an informative utterance ("Say enough") and one that mandates low speaker effort ("Don't say too much") (Horn, 1984; Horn, 2006). In Sperber and Wilson's relevance theory, the relevance of a stimulus such as an utterance is higher the more cognitive effects it has for the *hearer*, and lower the more effort it requires to process (to derive those effects) (1986).

Theories of classical or idealised rationality do not take into account the limitations imposed on humans by our limited time and resources for representing and processing information: the 'finitary predicament' (Cherniak, 1981). The most implausible type of theory would assume that processing and information search are costless; so the very best solution to any problem is found, taking into account all relevant information weighted appropriately, no matter how implausible it is that this could be achieved.

Pragmatic theories that propose a balance between results and the effort expended avoid the trap of assuming that rational agents operate without any cost considerations, but this is not enough to ensure that they make realistic assumptions. A variant of classical rationality explicitly allows for costs as well as benefits – for example in decision theory and game theory (Simon, 1983, pp. 12–17) – but is still an impossible idealization since it assumes that "the decision maker contemplates, in one comprehensive view, everything that lies before him" (Simon, 1983, p. 13) and chooses the best option, taking costs into account, that is, the one that best balances benefits and costs. This is 'optimisation' in Simon's terms.

Even the weaker idea that optimisation provides a standard to aim at should be treated with caution, in my opinion. For example, Kasher's version of the rationality constraints on utterance production requires a 'rationally ideal' speaker to optimize in this sense: "given a desired end that can be obtained only by some speech act, a rationally ideal speaker opts for a speech act that, to the best of one's belief, attains that end most effectively and at least cost, *ceteris paribus*". (Kasher, no date). Depending on whose cost is to be minimised this either amounts to choosing a maximally efficient utterance, i.e. one that conveys the speaker's intended meaning at least cost to herself; or to the speaker minimizing the effort required by a hearer, so that, given a particular utterance, a hearer can simply look for the interpretation that provides most information for the least effort.

Real speakers and hearers are not rationally ideal, of course, and those who think that rationally ideal speakers maximize effects achieved for effort expended do not necessarily suppose that real speakers do. In my opinion one can go further than that: it is implausible that speakers or hearers even aim to maximize in this way, or that hearers proceed on the assumption that speakers do.

In arguing that this is not the way communication works, I draw on various strands of work in psychology and philosophy which converge on the claim that classical, ideal rationality is unattainable by human beings in principle and in practice. Christopher Cherniak has argued that the (correct) concept of rationality is much more minimal than classical theories propose (Cherniak, 1981; Cherniak, 1986). In practice, it is clear that we are finite beings and we work within the restrictions that imposes, as argued by Herbert Simon (e.g.Simon, 1955; Simon, 1956; Simon, 1957a; Simon, 1969) and more recently Gerd Gigerenzer and colleagues (e.g. Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999).

On the short timescales involved in quick inferences or decisions, including most utterance production and understanding, it is implausible that we act as though we take all information into account ('optimize' in Simon's sense), finding the best possible balance of cost and payoff, a global maximum. Rather we use procedures that aim to find solutions that are good enough by searching and stopping once expectations are satisfied. In special cases we follow procedures that aim at *local* maxima, trying to find solutions that are the best within the compass of a limited search. Basing my account on Sperber and Wilson's relevance-theoretic comprehension procedure, I will argue that utterance understanding is one of these special cases, an expectation-based search which stops only when its quite specific aspiration-level is attained.

Heuristics

Generally, instead of behaving as though we weighed up all of the options and information, as classical agents do, we use heuristics that allow us to ignore large amounts of information by making use of properties specific to the task.

The way that we catch balls is an example. (McLeod & Dienes, 1996; Gigerenzer, 2001, pp. 3007–3008). Successful catching depends on environmental regularities that we become attuned to. The gravitational field where we live, near to the surface of the earth, is nearly uniform, so the acceleration of any object due to gravity, *g*, is nearly constant. This means that a thrown object will move in a certain kind of nearly parabolic path. (It would be exactly parabolic without wind resistance, but this is non-negligible for thrown balls (Brancazio, 1985)). It happens that this means that one can catch a ball by moving backwards or forwards so as to keep the angle between the ball and one's eyeline increasing at a certain rate, and thereby ensuring that it will stay between 0° and 90° until the ball is within reach (McLeod & Dienes, 1996)[8].

The procedure is a heuristic. It works under the right conditions, getting the fielder to the same place as the ball, as long as he can run fast enough.[9] It would not work reliably under other conditions, such as non-constant *g* or with a self-propelled object. It is only applicable to one environmental problem, catching objects, and is not applicable outside of this domain. This kind of problem specificity or domain specificity[10] is a property of heuristics in the

[8] To be more precise, the fielder keeps the second derivative of the tangent of the angle equal to zero: $d_2(\tan \alpha)/dt^2 = 0$; equivalently he keeps $d(\tan \alpha)/dt$ constant (McLeod & Dienes, 1996).

[9] This heuristic only solves the part of the problem of catching concerned with how far along its trajectory a ball will be at catching height. A catcher generally has to move left or right as well as towards or away from the point of projection. A separate procedure takes care of getting into the correct position laterally (McLeod & Dienes, 1996, p. 532). A combination of the two procedures is used in the general case (McLeod, Reed, & Dienes, 2001; McLeod, Reed, & Dienes, 2003; McLeod, Reed, & Dienes, 2006). A further procedure is used as a fine adjustment at the last moment. The fielder stretches out his arms forward or above his head, diving forward or jumping upwards as necessary (McLeod & Dienes, 1996, p. 537).

[10] Problem specificity and domain specificity are not generally the same. A heuristic may be useful for similarly structured problems in different domains. The recognition heuristic, for

sense that I use the word in this thesis[11]. The procedure requires only a small amount of mental resources and information, using only the angle of gaze as a cue. It is fast and frugal, in Gigerenzer's terminology.

In comparison, a rigorous calculation of the intersection of the ball's trajectory with the ground would require more mental resources and more information from the environment: projection angle, initial speed, and wind-resistance at least (McLeod & Dienes, 1996, p. 531; Gigerenzer, 2001, pp. 3007–3008). For a fully accurate calculation, the spin of the ball and the humidity and wind-speed would also need to be ascertained and taken into account. In contrast with the heuristic, a truly rigorous calculation using the initial velocity and the law of gravity to calculate the trajectory, with modifications for wind-resistance, spin and other factors, would work under any conditions, but at the cost of vastly increased effort and information required.

The reason why a heuristic only reliably works under certain circumstances is because assumptions about the structure of the task environment that are not true in all domains are built in, rather than explicitly given as premises or parameters. (For example, the ball-catching heuristic has built-in the assumption that acceleration is close to constant throughout the flight.) This contrasts with algorithmic procedures such as arithmetic, or truth-table proofs in classical logic, which guarantee correct answers; and with Bayesian, decision-theoretic, and game-theoretic accounts of decision-making where the decisions to be made are assumed to be those that would be reached if all available information were considered and taken into account. In such cases, domain-specific assumptions must be put in place in order to solve problems in the relevant domain, and since the mechanism is domain-general overall, these assumptions must be explicitly included.

There is a further difference between heuristics and idealised classical rationality. For heuristics – and for bounded rationality generally – search is

example, works well when the sole cue, that is, whether an item is known or unknown, is correlated with the criterion, whether the problem is determining the larger of a pair of cities or picking stocks for a portfolio. (Goldstein & Gigerenzer, 1999; Borges, Goldstein, Ortmann, & Gigerenzer, 1999; Goldstein & Gigerenzer, 2002; Todd & Gigerenzer, 2003, pp. 149–150, 155–157)

[11] I explain this choice and look at other uses of the term in chapter 3.

crucial. The catching heuristic is a good example. It takes the catcher to the correct position at the correct time to catch the ball, but the catcher does not explicitly[12] calculate where the ball will land (McLeod & Dienes, 1996, p. 539)[13].

If humans were ideally rational, with no time-constraints or resource limitations on calculation or information gathering, they could make all judgments and decisions using domain-general procedures with situation-specific information explicitly fed in. Under time and resource limitations, however, it makes sense to have procedures that do not allow for the vast range of possibilities but work reliably in a small corner of human experience. Some such procedures will be highly innately-specified, others will depend more on experience of the environmental regularities in the relevant domain. A collection of such procedures all applicable to one domain might be seen as a mental organ or module.

In their recent work, Sperber and Wilson view the utterance comprehension system as a dedicated module of this type.[14] This module includes the relevance-theoretic comprehension procedure, a fast and frugal heuristic (Sperber & Wilson, 1986, p. 45; Sperber & Wilson, 2002, p. 9; Sperber & Wilson, 2004, p. 624)[15], which makes use of regularities that are specific to inferential communication about a speaker's intentions. One centrally important regularity is described in the *communicative principle of relevance*, which licences the *presumption of optimal relevance* (both from Sperber & Wilson, 1995):

[12] The point is not that the fielder is not conscious of a calculation of the position where the ball will land, but that no such calculation is made. That is, the relevant pieces of information are not mentally represented nor mentally manipulated in the way that such a calculation requires.

[13] If fielders did this, they could run to the position where the ball lands and wait for it. They do not do this, instead running through the catching position as the ball reaches it.

[14] The question of modularity is not central to the concerns of this thesis. Pragmatic inference, whether or not it is carried out by a module, proceeds fast without consulting all potentially relevant information. See chapter 5.

[15] Indeed in a recent overview, Sperber and Wilson use the term 'relevance-theoretic comprehension heuristic' (2005, p. 360, and thereafter) to the exclusion of the previous 'relevance-theoretic comprehension procedure'. I keep the older formulation here, since part of what I am discussing is whether the procedure is in fact a heuristic.

(1) *Communicative principle of relevance:*

Every ostensive stimulus conveys a presumption of its own optimal relevance. (Sperber & Wilson, 1995, p. 260)

(2) *Presumption of optimal relevance:*

a) The ostensive stimulus is relevant enough for it to be worth the addressee's effort to process it.

b) The ostensive stimulus is the most relevant one compatible with the speaker's abilities and preferences. (Sperber & Wilson, 1995, p. 270)

This presumption is a precise proposal about what it is that a hearer is rationally justified in expecting from any utterance intended for him. If these expectations are justified, then, faced with an utterance, it makes sense to look for an interpretation that satisfies them. Sperber and Wilson propose that the relevance-theoretic comprehension procedure does just this. In chapter 5 I look at the properties of searches that exploit these regularities.

1.1.5 SUMMARY

This thesis combines three key elements. The first is the broadly Gricean view of utterance production as intentional action intended to induce the hearer to recognise the intention behind the action, and a corresponding view of utterance understanding as a process of grasping the relevant speaker intentions. This makes utterance interpretation a form of inference to the best explanation and utterance production a matter of devising clues that will be interpreted correctly. Thus, as Grice put it, "the idea [is] that the use of language is one among a range of forms of rational activity" (Grice, 1989b, p. 341).

The second is a realistic version of a traditional view of human rationality. I argue that realism about human abilities requires that we view rationality as bounded and mostly implemented by heuristics (chapter 3). My view is traditional, though, in that I argue that rationality is centrally the ability to reason and that reasoning involves ability with truth-preserving logical operations (chapter 2). I draw on Grice's work here too, in reconciling a traditional view of this type with the reality of fast and frugal reasoning. His position was that there is "a 'hard way' of making inferential moves; [a] laborious, step-by-step procedure [which] consumes time and energy... .A substitute for the hard way, the quick way, ... made possible by habituation and intention, is [also] available to us" (Grice, 2001, p. 17).

A third key element, not described so far, is a view of the mind/brain as a device which processes information (i.e. a view congruent with modern psychology and modern linguistics). It is this commitment to understanding cognition as computation over mental representations that makes it clear that, to avoid computational explosion, heuristics must be used in cognition.

1.1.6 CARTESIAN THEORIES OF COGNITION

Representational-computational theories are a product of what Chomsky calls the second cognitive revolution, the renaissance of mentalist and nativist explanations in linguistics and psychology in the late 1950s and 1960s, reviving Cartesian and other rationalist views in the context of a greater understanding of computation. This view of psychology treats it as a branch of natural science: specifically, the branch which tries to explain thought and behaviour in terms of states of the mind/brain (Chomsky, 1991a, pp. 4–5). This thesis is intended in that spirit as a contribution to the explanation of how we understand and produce utterances in terms of mental states, specifically mental representations. Later I discuss mentalistic approaches to central, conceptual cognition (in chapter 2), and to utterance understanding (in chapter 4). In this introduction I make only brief remarks on the general project.

One important part of the second cognitive revolution has been the re-emergence of Cartesian representational theories of perception. In contrast to empiricist theories which treat the mind as a passive recipient in perception, in the Cartesian picture the mind generates a representation of objects on the basis of perceptual stimuli, but going beyond them (Chomsky, 1991a, p. 14).

> ... the eye scans a surface, or a blind man taps it with a stick... The mind then uses this sequence of impressions to construct the representation of a cube or a triangle or a person, employing its own resources. (Chomsky, 1991a, p. 14)

The generative grammar research programme in linguistics adopts the same approach: "the mind produces the representation of a presented expression, making use of the I-language and of course much else." (Chomsky, 1991a, p. 14)

It is an assumption in this thesis (as in relevance theory) that utterance interpretation is to be accounted for in a similar way, as the generation of a mental representation of the meaning of an utterance as a response to stimuli relating to that utterance. I will argue, though, that there is a crucial difference between the pragmatic process and perceptual processes. As I have said, I agree with Sperber and Wilson that utterance understanding is an inferential process, in contrast to linguistic parsing or visual processing, which are non-inferential processes[16], effectively very complex reflexes. In addition, I argue that utterance interpretation is a reasoning process in that the representation generated in successful utterance interpretation is not only caused by but also rationally justified by the utterance, in that the utterance (and other clues) provide good evidence for the interpretation. This relies on a further key difference between pragmatic processes and perceptual processes. Pragmatic processing is a central process rather than a peripheral, perceptual one, in that it takes propositional input, rather than, as perceptual processes do, input directly from transducers connected to sense organs. (I return to these points in chapter 4).

1.1.7 ASSUMPTIONS

In the remainder of this introductory chapter I comment on some issues which concern alternatives to assumptions which I make. One crucial assumption is that communication involves inferences about speaker's intentions. In section 1.3 I briefly consider work in psychology on the development of the ability to reason about other's mental states, and comment on anti-inferentialist theories of communication.

First, I consider the relationship between reasonableness and rationality. The legal conception of what it is to be reasonable suggests that the views that

[16] Linguistic parsing is a decoding process; visual processing is non-inferential but not (strictly) a decoding process, since visual cues are natural signs rather than a code.

rationality is bounded has deep roots, as Cherniak pointed out. The contrast with reasonableness helps to create room for a view of rationality as what an agent is *able* to do, rather than what he actually does. (I develop this view of rationality as reasoning ability further in chapter 2.)

1.2 REASONABLENESS AND RATIONALITY

While this thesis addresses rationality, not reasonableness, some of what matters for understanding rationality is arguably to be learned from the ordinary notion of reasonableness and derivative concepts such as *reasonable person*. There are competing conceptions of *reasonable* in the law, where it is an crucial notion, as well as in philosophy, and competing ideas of how it relates to rationality. I explore some of these links in this section, then set aside the concept *reasonable* for the remainder of this thesis.

In English law one conception of *reasonableness*, close to the ordinary meaning, is something like *proportionate*. A reasonable person in law, for example, is one who takes account of such possibilities as could be expected to occur, and exercises due care towards possible occurrences which are unlikely but not so improbable or out of the ordinary as to be unforeseeable (Cherniak, 1986, pp. 101-102).

This conception of *reasonable* coexists uneasily in law with another meaning closer to that of *rational*:

> It is extremely difficult to state what lawyers mean when they speak of 'reasonableness'. In part the expression refers to ordinary ideas of natural law or natural justice, in part to logical thought, working upon the basis of the rules of law. (Salmond, 1947, Jurisprudence, quoted in Garner & Black, 2004, p. 1293)

Whether *reasonable* relates to logical thought or to something closer to common-sense is an important matter in law, not least because 'beyond a reasonable doubt' is the standard a jury must use to decide whether a defendant in a criminal case is guilty (Garner & Black, 2004, p. 1293). One controversial suggestion is that a reasonable doubt is – like a rational belief in philosophy – one for which there is a (good) reason.[17]

> In one sense the word [*reasonable*] describes the proper use of the reasoning power, and in another it is no more than a word of assessment. ... Lawyers say a reasonable doubt, meaning a substantial one; the Court of Appeal has frowned upon the description of a reasonable doubt as one for which reasons could be given." (Devlin, 1979, in The Judge 134, quoted in Garner & Black, 2004, p. 1293)

The meaning of *reasonable* as something close to 'in proportion, as common sense would have it' is dominant. 'Reasonable' has a similar meaning in the concept *reasonable person*, derived from ordinary usage and folk psychology and embodied in English common law. This concept is a key notion in negligence cases. A person cannot be expected to foresee all consequences of his actions. Tort law[18] recognises this by postulating a hypothetical reasonable person used as a yardstick. One may be negligent if one fails to consider the possibilities that would be considered by a reasonable person and to act appropriately. One cannot be held negligent for failing to take action that a reasonable person would not take.[19] (Cherniak, 1986, pp. 101–102)

As Cherniak points out, there is a parallel with theoretical discussions of agents who are rational but less than classically so. What it is rational for an agent to do or to believe depends on which factors can realistically be taken into account, and, because agents have finite resources, it is not possible that every confound to the truth of a belief or to the desirability of an action is taken into account by an agent. In contrast, classical rationality requires that an agent consider all relevant factors. It seems that the legal conception *reas-*

[17] The issue is actually even more complicated. The word 'rational' is sometimes used in cases where reasonable – in the sense close to proportionate – is generally used. Thus, for example, "'rational doubt' means the same as 'reasonable doubt'" (Garner & Black, 2004, pp. 1290, 1293).

[18] Tort law is the area of law concerned with breaches of obligations that people owe to each other, except for contractual obligations (Garner & Black, 2004, pp. 1290, 1526).

[19] Less is required of those with 'diminished responsibility', children, for example. Conversely, the standard required is higher for experts such as doctors, engineers and lawyers. (Cherniak, 1986, p. 103; Martin, 2002, p. 409)

onable person is aligned with the idea of a finite agent.[20] As Cherniak remarks, "standards on an agent's performance that fall short of unequivocal perfection are not just an unwieldy philosophical contrivance; they have been used traditionally as a core element of procedures in a domain of great practical importance." (Cherniak, 1986, p. 102) I return to classical rationality, bounded rationality and Cherniak's 'minimal rationality' in chapter 3.

In philosophy, it has been proposed that reasonable behaviour is appropriate and balanced behaviour. Reasonableness is often seen as going beyond rationality. Robert Audi's views are a good example. For Audi, "nothing reasonable fails to be rational; but a rational person, or stance, can surely fail to be reasonable." (Audi, 2001, p. 149) Reasonableness consists of rationality, which can be seen as the property of "conforming to logical and epistemic standards", plus something else: "the sort of thing one would expect of a rational person who is at least moderately thoughtful and balanced" (Audi, 2001, p. 149).

According to Audi, rationality is something like a capacity; the ability to do things logically. Having this capacity does not determine how it is used. To say that a person or stance is reasonable, on the other hand, is to say something about actual conduct.

A second difference is that "reasonableness requires a greater responsiveness to reasons than mere rationality" (Audi, 2001, p. 149). This seems to mean that a reasonable person must be able to tell good reasons from bad reasons or the absence of reasons: "being a reasonable person requires a measure of good judgement and is incompatible with pervasively bad judgement" (Audi, 2001, p. 150). Thus reasonable people act on good judgement *more* than merely rational people do and they act on a whim (i.e. for no particular reason) *less* than merely rational people.

A related requirement is that "reasonable people are to some degree self-critical". (Audi, 2001, p. 150) This is congruent with the requirement of "responsiveness to reasons" since being self-critical is plausibly helpful in appreciating whether one has good reasons for one's attitudes.

[20] The remarks here about the conception of 'reasonable' in law are only intended to demonstrate this point, and are far from a complete survey. For example, there is a tort of nuisance, distinct from negligence, and there the term is used somewhat differently (Jones, 2002, p. 333).

It follows from Audi's conception of rationality as a capacity that irrationality excuses – or mitigates – foolish or bad behaviour that occurs as a result of it, as for example in children; unreasonableness does not, since an unreasonable person is one who has the ability to fit beliefs and actions to reasons but will not.

I think Audi's way of making the distinction draws out some of the implications of normal usage. I agree in this thesis with his conception of rationality as a capacity. I would say that it is a kind of dispositional property of a mental system: what it is capable of doing with incoming and stored information in virtue of its internal structure. His conception of reasonableness as both possessing these capabilities and using them correctly, so that reasons guide beliefs and behaviour suggests that it is a doubly normative concept. Rationality is the ability to deal correctly with beliefs etc.; reasonableness is the correct use of this ability.

There is some similarity with Grice's division of rationality into a basic reasoning capacity, flat rationality, and higher levels of reasoning ability, variable rationality (Grice, 2001, pp. 28–36), which I discuss in chapter 2. A difference is that Grice does not mean by variable rationality the correct use of the abilities making up flat rationality; rather possession of a higher degree of variable rationality is the possession of greater reasoning ability, including some capabilities useful for reasoning but not essential to it.

Grice also comments briefly on reasonableness (2001, pp. 23–25), taking the view that 'reasonable' and 'rational' refer to two different general qualities of reason (Grice, 2001, p. 23). He draws on a distinction made by Aristotle in the Nicomachean Ethics. Aristotle distinguishes between parts of the soul which possess reason intrinsically – the parts that do the reasoning – and non-reasoning parts of the soul which possess reason extrinsically insofar as they are governed by the principles of reason.[21] Grice suggests mapping this

[21] Views along these lines became a commonplace in the classical world, forming, for example, a basis for Stoicism. Cicero reflects Stoic values in the famous line, "Reason should direct and appetite obey." Plato had also previously made a distinction between the rational part of the soul, which may override the (non-rational) appetites:

> ... the [principle of the soul] with which a man reasons, we may call the rational principle of the soul, the other, with which he loves and hungers and thirsts and feels the flutterings of any other desire, may be termed the irrational or appetitive, the ally of sundry pleasures and satisfactions. (Plato, 1991, Book IV)

distinction onto the distinction between rationality and reasonableness so that in behaviour, rationality is possession or display of "the capacity to reach principles or precepts relating to conduct" (Grice, 2001, p. 24) and to be reasonable is "to be free from interference, on the part of desire or impulse, in one's following such principles or precepts." (Grice, 2001, pp. 24–25). In this view reason is a kind of regulation. This has the advantage that we can say that behaviour that is not according to reason (unreasonable behaviour) need not be due to irrationality. Rather, what is being regulated, the parts of the soul (mind) that are not in themselves reasonable, may have got out of hand. This should allow a simpler theory of rationality than one which must account for all lapses as failures of reasoning (Grice, 2001, p. 25), just as the competence-performance distinction makes a theory of grammar possible by removing the necessity for the grammar to generate the ill-formed utterances that are made when tired, drunk or confused.[22]

There are examples which seem to challenge Grice's view. In an episode of the Simpsons, Homer, trying to steal coke and sweets, gets his hands stuck in vending machines. Help is summoned but the hands cannot be freed and he is told that his arms will have to be amputated. Just in time, it is noticed that Homer is holding on to a can inside the machine.[23] I think we would call Homer's behaviour irrational rather than unreasonable (or perhaps both irrational and unreasonable). It is clear, though, that on Grice's definitions Homer is behaving rationally but unreasonably. Homer can work out the consequences of his actions (although he hopes his arms will grow back afterwards), but is unable to act appropriately precisely because he is overcome with desire.

Perhaps Homer's actions seem irrational because his actions are not well suited to achieve his desires. This would accord with a well-known version of

[22] Grice does not draw the parallel with generative grammar. The case of rationality is complicated by the fact that we can say that a *person* is rational as well as a belief, whereas only sentences are grammatical or otherwise.

[23] *Marge on the lam*, episode 1F03, season five, first aired 5th November 1993. The dialogue in the scene is as follows: FIREMAN: Homer, this... this is never easy to say. I'm going to have to... saw your arms off. [*brandishes a circular saw*] HOMER: [*plaintive*] They'll grow back, right? FIREMAN: Oh, er, yeah. HOMER: Whew! SECOND FIREMAN: Are you just holding on to the can? HOMER: Your point being? (Adapted from http://www.snpp.com/episodes/1F03.html)

the contrast between rationality and reasonableness made by the political philosopher John Rawls, who draws on work by WM Sibley (Sibley, 1953).

> Knowing people are rational we do not know the ends they will pursue, only that they will pursue them intelligently.
>
> Knowing that people are reasonable where others are concerned, we know that they are willing to govern their conduct by a principle from which they and others can reason in common. (Rawls, 1993, p. 49)

What is to be taken away from this discussion? Two points seem important. First, as Cherniak suggests, there are indications in the ordinary use of the word 'reasonable' and its precipitate over time in the terminology of common law that folk psychology regards people as bounded agents, capable of paying attention to some reasons for actions and beliefs but not all. Secondly, some philosophical discussion of the difference between reasonableness and rationality suggests that rationality should be seen as a capacity or faculty that may or may not be manifested in any particular judgement or action. This opens the way to a simpler, competence theory of rationality and is perhaps a prerequisite for any realistic attempt at such a theory.

1.3 ALTERNATIVES TO INFERENTIAL-INTENTIONAL THEORIES

1.3.1 CODING AND INFERENCE

Grice's major achievement in the field of communication was to show that what a speaker meant by an utterance must be inferred. His theory of meaning, and other theories which follow him in arguing that hearers infer speakers' intentions (I will call them *inferential-intentional* theories), are therefore in implicit opposition to an older theory, the code model of communication (Sperber & Wilson, 1986, pp. 2–21, 44–6; Sperber, 1994).

According to the code model, communication involves the transmission of a meaning – the message – by means of language. The idea is that the speaker encodes and transmits her meaning as a linguistic signal, which the hearer then decodes (Sperber & Wilson, 1986, pp. 4–5; Sperber, 1994). The 'message'/'signal' terminology is from information theory (e.g. Shannon & Weaver, 1949, p. 3).

There are two fundamental differences between the code model and the inferential model. First, the relationship between the signal and what it encodes is arbitrary in the sense that it does not provide evidence for the message, absent the code. As Sperber says, "just as the letter 'm' does not logically follow from two long beeps [its symbol in Morse code], the meaning of a sentence does not logically follow from its sound" (1994, pp. 181–182). In contrast, the inferential model treats utterances, their features, how they are made, and that they are made, as clues to the intended meaning. The meaning can be worked out on the basis of the utterance, together with appropriate background assumptions, and follows logically (although non-demonstratively) from the fact that *that* utterance has been made (by a certain speaker, in certain circumstances, etc.).

Secondly, a coding/decoding process will lead to perfect transmission of the message if certain conditions are met. That is, there will be perfect transmission if the code is shared, encoding and decoding are carried out successfully, and the signal is not distorted by noise or interrupted. There are in principle no strong guarantees of that sort for inferential processes: the hearer may work out what the speaker's intended meaning was or he may not.

A third difference rests on these. As Sperber points out, there is no room for creativity in encoding or decoding: strictly speaking, applying a code creatively is applying it wrongly (Sperber, 1994). In contrast, working out what logically follows from an utterance is a creative process. It involves postulating (or generating) a conclusion, assessing whether the utterance (or the way it is made, or the fact that it has been made) supports it, that is, whether the conclusion follows from the utterance together with other assumptions, and whether those other assumptions are plausible or at least not unreasonable.

I do not provide arguments here that a view of communication purely in terms of coding and decoding is untenable. In my opinion, Grice's work establishes this, setting the bar higher for anti-inferential views of communication. Any theory of communication has to give some explanation of how hearers work out parts of speaker meaning that are not part of the stable, encoded meaning of the linguistic items used.

A danger for anti-inferentialist models is that they will reduce to the code model and will be unable to account for disambiguation and reference assignment as well as apparently more complex phenomena such as implicature and modulation of word meaning. Alternatively, they may slip towards redescription of the problem, noting that linguistic items are often used to convey meanings beyond, or at variance with, their fixed meanings, but failing to give an account of the processes involved.

1.3.2 ANTI-INTENTIONALISM AND ANTI-INFERENTIALISM

A number of theorists espouse views of communication which are *prima facie* distinct both from Gricean inferential-intentional pragmatics and from the code model, among them the philosophers Ruth Millikan (1984; 2005) and Tyler Burge (1993) and recently linguist Richard Breheny (2006).[24] These theories are all built on an intuition that some aspects of normal conversation are less complex and more direct than inferential-intentional theories propose. They are avowedly non-intentional, partly or fully, in that they propose, *contra* Grice, that communication need not involve a hearer's recovery of speaker's intentions.

Millikan claims that in the normal flow of conversation, utterance understanding is essentially a form of perception, unmediated by reasoning about the speaker. Breheny (2006) proposes that some communication – 'basic communication' – can take place in the absence of thoughts about a speaker's thoughts. According to this view, speaker's meaning and hearer's understanding can be coordinated purely by attention to shared situations, in basic cases

[24] I do not deal in this chapter with François Recanati's view that 'primary processing' is a brute, non-inferential process. Recanati (2002a) divides pragmatic processing into two parts, primary and secondary, claiming that only secondary pragmatics involves inferential recovery of intentions. He argues that primary pragmatic processing, the understanding of what is said, involves only non-inferential processes: "primary pragmatic processes ... need not involve an inference from premises concerning what the speaker can possibly intend by his utterance. Indeed, they need not involve any inference at all: communication, I argue, is as direct as perception." (Recanati, 2002b, p. 105)

But Recanati thinks that what the hearer recovers by primary processes is what is said in a Gricean sense, that is, a hypothesis about part of speaker meaning. I therefore reserve consideration of Recanati's theory to chapter 4 below, where I consider the nature of inferences in Gricean communication.

at least. Breheny claims that this renders explicable the communicative ability of young children who are incapable of reasoning about each other's states of mind.

Both Millikan's theory and Breheny's basic communication are intended as radical alternatives to intentional theories of pragmatics[25] and part of the stated motivation in both cases is dissatisfaction with the explanatory status of a broadly Gricean theory of communication.

1.3.3 EXPLANATORY STATUS OF GRICEAN PRAGMATICS

The worry about Gricean explanations is that a crucial aspect is reasoning of some complexity about speaker's intentions, and the status of such explanations is in doubt if, as it seems, hearers do not explicitly reason in this way. As Millikan says, "Mere behaviors don't explain anything. Only their underlying causes are explanatory." (2005, pp. 203, note 6.) (See also Breheny, 2006, pp. 101–102, for similar concerns.)

This point about explanation is central to this thesis. I will argue that inferential-intentional reasoning is explanatory, even when the processing does not mirror the argument. A major purpose of this thesis is to spell out how a broadly Gricean account is explanatory without postulating that processing is so complex as to make implausible demands, not just on young children, but also on finite agents in general. Anticipating the discussion somewhat, I want to make two points about non-Gricean theories.

The first is that such theories may not be entirely non-inferential (contrary to the theorist's intentions) if the output of processing is represented as speaker's meaning (or some component of it such as 'what is said'). A fast, automatic process for arriving at speaker's meaning is, in my view, a fast, automatic process for performing a certain kind of inference.

A process of this kind might automatically take into account such useful cues as direction of gaze for fixing reference. It might also have a rule that

[25] Both think that *some* communication requires hearers to make inferences about speaker's intention. For Millikan this only occurs when the normal flow of conversation is disturbed in some way. Breheny proposes basic communication to account for the communicative abilities of children he thinks too young to carry out inferences about communicative intentions. It is not clear to me what he thinks is the division of labour in adult hearers between basic communication and inferences about communicative intentions.

makes the search terminate as soon as a coherent interpretation is found. Such factors would make a process non-algorithmic, that is, heuristic, but they do not make it non-inferential. A procedure for finding meaning quickly might not explicitly represent the beliefs of the speaker, but if it works out speaker meaning on the basis of reliable (albeit fallible) cues to a speaker's mental state such as her direction of gaze, and it builds in some way of rejecting a trial interpretation as unsatisfactory then it is inferential.

In fact, I would argue that this is just how we should expect a very basic heuristic for inferring a speaker's meaning to work – such as, presumably, the comprehension procedure employed by young children. It should pick up on clues that are offered by modules that operate from infancy, such as gaze detection. If it is to be as computationally simple as possible, we should expect it to accept any solution that seems good enough, that is, to satisfice. (I discuss satisficing heuristics in chapter 3). We would expect it not to perform elaborate checks on the adequacy of the solution found. Thus for example if the speaker says "He's spiny", a young child might simply assign as the referent of the pronoun "he" whatever the speaker seems to be attending to, as long as it is plausibly semantically countable, singular and male. (This account receives some support from evidence that early vocabulary acquisition makes use of gaze detection. For example Paul Bloom thinks that the child automatically checks speaker's direction of gaze before assigning reference (Bloom, 2000, ch.3).)

Sperber suggests that young children in the first developmental stage of pragmatic ability – which he calls *naïve optimism* – accept the first interpretation that occurs to them that meets their expectations of relevance, that is, the first one that seems to deliver enough cognitive effects for the effort put in. (Sperber, 1994; see also Sperber & Wilson, 1987a where this kind of strategy was first suggested). This strategy is very simple, but inferential nonetheless.

A distinct view is that the hearer simply takes as correct the first interpretation that occurs to him. In this kind of theory, the interpretation is fixed by the facts about accessibility: for example, the salience of a referent in the context. In chapter 4 I discuss Recanati's view that the explicit meaning of utterances ('what is said') is derived this way by hearers.

A vital point for an inferential theory of speaker meaning is that the result of the procedure or procedures used is represented as speaker meaning[26], so that utterance interpretation is a matter of arriving at a hypothesis that meets certain standards about the speaker's meaning. By this criterion Millikan's theory seems to be genuinely non-inferential. She claims that the linguistic items used in utterances cause beliefs in the hearer directly and that the beliefs caused are about whatever the sentence or sentences used are about, not about the speaker's meaning. This, as far as it goes, is a non-Gricean account.

1.3.4 MILLIKAN'S PERCEPTION THEORY OF UTTERANCE COMPREHENSION

Millikan's view that utterance interpretation does not involve inference about intentions, nor indeed any thoughts about speakers' intentions, being more like perception than reasoning, is a long-standing alternative to the Gricean view of utterance interpretation as abductive inference about the speaker's intentions (Millikan, 1984; a useful recent summary is in Millikan, 2005). According to Millikan, "Speech is a form of direct perception of whatever speech is about. Interpreting speech does not require making any inference or having any beliefs ... about speaker's intentions" (Millikan, 1984, p. 62). This makes the distinction clear, although it is worth noting that the difference between this view and inferential-intentional theories is less than one might suppose, since Millikan's opinion is that perception is "itself not all that direct" (Jary, 2005, p. 93). Perception fills in gaps. Millikan gives the example of seeing part of a cat in long grass. A similar point was made by Hume:

> Suppose I see the legs and thighs of a person in motion, while some interpos'd object conceals the rest of his body. Here 'tis certain, the imagination spreads out the whole figure. I give him a head and shoulders, and breast and neck. These members I conceive and believe him to be possess'd of. Nothing can be more evident, than that this whole operation is perform'd by the thought or imagination alone. (Hume, 2003, p. 445)

[26] Deirdre Wilson (p.c., 11/2006) pointed out to me the importance of this issue to the question of whether a theory is inferential.

In chapter 5, I argue that perceptual processing is not inferential (see also Wilson, 2005, pp. 303, footnote 1; Sperber & Wilson, 1986, pp. 12–13), since the input is not conceptual. That discussion is about the process that takes as input activations on the retina, or the analogue representations transduced from them, and produces as output a three-dimensional model of the scene. Millikan and Hume's examples go further than this, and it is possible that some genuine inference is involved in concluding from glimpses of cat ears that there is a cat present, as there is in inferring the same thing from a miaow. If the claim is that arriving at speaker meaning is only as direct as this I would not necessarily disagree. However, Millikan means something much stronger: that the result of processing an utterance is knowledge about the world, unmediated by thoughts about the speaker's mental states.

There are two components to Millikan's theory: the claim that utterance understanding is direct perception, and a separate theoretical framework for language which treats individual constructions and lexical items as having functions.

Millikan thinks that linguistic items and linguistic forms (she uses the term 'linguistic device' to cover both) have purposes by virtue of which they continue to exist. For her, linguistic devices may be lexical items, surface syntactic constructions, phonological items such as a particular pattern of stress or intonation and even orthographic elements, such as punctuation systems: essentially all "significant surface elements that a natural spoken or written language may contain" (Millikan, 1984, p. 3). For Millikan, the purposes of linguistic devices involve direct modification of the thoughts of the hearer. For example, the word 'elephant' has the purpose of evoking thoughts of elephants (Millikan, 2005, p. 191); and indicative sentences have the purpose of "effect[ing] production of a true belief having whatever propositional content the various other aspects of the sentence are designed to impart." (Millikan, 2005, p. 190).

This account is grounded in a claim that the kind of purpose that linguistic devices have is their evolutionary *stabilising direct proper function*. (Millikan, 1993, gives definitions of these terms.) What Millikan means by this is that serving a particular purpose is what keeps a linguistic device in being. The stabilizing direct proper function of a linguistic device is the production of its conventional meaning in hearers (by conventional meaning, Millikan means something similar to Grice's 'timeless meaning' (1968), or 'linguistic meaning' in Sperber and Wilson's work (1986) and in this thesis).

Speakers have purposes too, which, according to Millikan, may differ from the purposes of linguistic items. This is Millikan's characterisation of the difference between linguistic ('conventional') meaning and speaker meaning. A speaker using the word 'elephant' metaphorically is using it with a purpose different from its stabilizing function. Irony and other figures of speech are to be accounted for in the same way. In conversational implicatures, "what the speaker means either conflicts with the stabilizing function of the form or has some additional purpose beyond." (Millikan, 2005, p. 191)

I have three criticisms of Millikan's theory. First (and least important here), it is not clear what explanatory work Millikan's notions of convention and function do in linguistics. Unless it can be shown that the notion of function in her sense is useful in explaining linguistic data, it is hard to see what role it plays in theorising about language. (See Millikan, 2003; Chomsky, 2003, pp. 308–315.)

Secondly, Millikan's theory gives the hearer the task of determining with what purpose a linguistic device has been used on a particular occasion. Since there are very many purposes that a linguistic device might have (derived or direct), utterance interpretation in Millikan's theory is a matter of resolving massive ambiguity (Origgi & Sperber, 2000). But how, without reasoning about which meaning the speaker intended? Until there is an explanation, Millikan has not made the case that analogies with perception are any more than that.

The task of a theorist is not finding (for example) considerations that render it "not surprising that when someone calls that they are ready, one generally knows for what they are ready." (Millikan, 2005, p. 211) The task is to explain *how* the hearer works out what the speaker is saying she is ready for.

Thirdly, I also share Origgi and Sperber's suspicion that Millikan's model is a version of the code model, "in that it explains communication by the systematic pairing of linguistic stimuli and responses", even though, as they say, "the responses she envisages are closer to perception on one side, to action on the other side, than the more abstract responses envisaged by standard [code model] accounts." (Origgi & Sperber, 2000, p. 149)

There are two further important lines of argument against non-inferential and non-intentional theories of pragmatics. The first is that inferential-intentional pragmatic theories have made considerable progress. That progress tends to support the truth of their central assumption[27] – that utterance understanding involves the recovery of a particular intention of the speaker – particularly in the absence of competing research programmes in pragmatics.

The second point is that a motivation which Millikan and, particularly, Breheny give for non-intentional theories seems to me to be less compelling than they claim. They both draw the conclusion from the literature on 'theory of mind' or 'mindreading' that young children lack the ability to attribute and/or reason about other agents' mental states.

According to Millikan, in the normal flow of conversation at least, "there are many ways of grasping the content that the specific speaker intends to convey without employing a theory of mind" (Millikan, 2005, p. 187) – in other words, without needing to take into account any mental states of the speaker, such as the speaker's intentions or beliefs. This would be an advantage for Millikan's theory, if, as many psychologists have thought over the last twenty years or so, young children cannot fully grasp others' mental states. (Millikan raised this point in her comments on early relevance theory (Millikan, 1987, p. 726). Sperber and Wilson's response is at Sperber & Wilson, 1987a, p. 737.) However, I think that the evidence now available suggests that while young children are not able to discuss others' intentions, desires and beliefs, they do sometimes take them into account, particularly in communicative situations.

[27] Sperber and Origgi make this point:

> The whole of modern pragmatics is predicated on this assumption, and its findings are arguments in favour of it. Of course, this does not make the assumption right, but those who deny it, are, in effect, implying that pragmatics as currently pursued is a discipline without an object, somewhat like the study of humours in ancient medicine. Surely, the burden is on them to show how pragmatics fails, and what is a better alternative to explain comprehension. (Origgi & Sperber, 2000, p. 156)

1.3.5 PRAGMATICS AND MINDREADING

In the well-known *Sally-Anne* or *false belief* task, a participant and a doll called Sally see an object hidden in location *a*. Sally then leaves, and the object is moved by Anne, in sight of the participant, to location *b*. Sally comes back and the participant is asked where Sally will look for the object. Participants younger than about four years old (and many autistic participants) mostly say that Sally will look in location *b*, i.e. where the participant knows the object to be. From around 4 years old, participants generally say that Sally will look in location *a*. This has been taken to indicate that from this age, children's responses are based on a representation of another's mental representation, different from their own. The ability to infer and represent other's mental states (and act on that basis) is called *Theory of Mind*, or *mindreading*. (Wimmer & Perner, 1983; Baron-Cohen, Leslie, & Frith, 1985; Wellman, Cross, & Watson, 2001. See Bloom & German, 2000, pp. 2–3, for a sketch of the history of the task.)

Mindreading ability seems then to emerge around four years old, if passing the false belief task is used as the criterion. However, younger children are at least somewhat competent with aspects of utterance understanding and production (notwithstanding their limited linguistic and attentional abilities). On an inferential-intentional theory this requires attribution of intentions or beliefs. The anti-inferentialist conclusion is that normal communication, at least in basic form, cannot be dependent on the ability to make inferences about a speaker's mental states.

However, advocates of inferential-intentional theories can argue that it is highly plausible that 1) we are particularly good at reasoning about people's communicative and informative intentions, 2) an innate ability to reason about such intentions would be especially useful and might be expected to come online very early in children, perhaps before a more general ability to reason about agents' beliefs, desires and intentions in other domains. As Wilson (2005) writes, "there is good reason to think that pragmatic interpretation is not merely an application of general mind-reading abilities to a particular (communicative) domain" (p. 306–7). Children capable of communication might fail false belief tasks because 1) they have abilities for reasoning about agents' mental states, but it is harder for them to use them outside the communicative domain; or, 2) they have acquired the specific ability to reason about intentions involved in communication but have not yet acquired the more general ability required to pass the false belief task. A combination of the two explanations is also possible.

The mindreading and developmental pragmatic evidence do not at all rule out dedicated abilities for reasoning with intentions in communication. Origgi & Sperber (2000, p. 163) point out that attributing speaker meaning in a Gricean framework and passing the false belief task are quite different abilities which require different mental resources. Representing a speaker's meaning involves the ability to entertain a second-order metarepresentation (at least) of a specific type:

(3) "She intends

me to believe

that it is time to go home" (Sperber, 1994, p. 186)

On the other hand, to pass the false belief task a child must predict behaviour on the basis of the evaluation of another's belief, a first-order metarepresentation, as true or false. The metarepresentation is of the form:

(4) She believes

the cat is in the green box.

A failure to pass the false-belief task could be due to a) misevaluation of the belief in question, b) failure to predict behaviour following from the falsity of the belief, or c) failure to represent another's belief in the first place. Children lacking in any of the relevant abilities might still be capable of constructing the very specific type of second-order metarepresentations that are needed, according to inferential-intentional theorists, to represent speaker meaning.

In my opinion, opponents of inferential pragmatic theories should specify the kind of representation that they believe children come to have as a result of understanding an utterance. As noted above, Millikan has made the bold claim that the representation is a representation of whatever it is the sentence uttered is about: a hearer processing an utterance about a cat on a mat ends up with a belief about a cat and a mat and (generally) no mental representations involving the speaker or her intentions. If a theory of communication along these lines were tenable (and I do not believe it is, for the reasons I have given above), then non-inferentialists could claim that young children can communicate before passing the false-belief task because they are incapable of metarepresentation in any domain.

However it is worth considering whether children younger than four years old are indeed devoid of mindreading and metarepresentational abilities for general tasks or for communication.

Recent developments

While children less than about 4 years old do not pass the standard false-belief task, there is evidence that young children can and sometimes do take others' mental states into account. Opinions differ as to how capable children are of reasoning about others' mental states. But evidence is mounting that well before passing the false belief task, and perhaps from as early as can be tested, children have expectations about others' mental states and act on that basis. Summarising this work recently, Enfield and Levinson write that "A number of researchers ... believe that children grasp the nature of the other as an intentional agent from about nine months" (2006, p. 16).

Bloom and German (2000) contrast 3-year-old children with older autistic individuals. Autistic children may really lack theory of mind, they argue, but they are very different from three-year-old typically-developing children, in that "Normal 3-year-olds are far superior with regard to communicative and linguistic skills, the ability to pretend and understand the pretence of others, and the ability to engage in, understand and manipulate the actions of others." (Bloom & German, 2000, p. B29) They conclude that three year olds fail the false belief task because the task is too demanding, because they do not grasp *false* belief, or both,

> But they surely have a 'theory of mind', in the general sense of having a sophisticated ability to reason about the mental states; this is precisely

> why they differ from autistic individuals in the social, communicative and linguistic domains. (Bloom & German, 2000, p. B29)

There is direct evidence that young children do keep track of others' false beliefs. Southgate, Senju, and Csibra (2007) found evidence that 2-year-olds "correctly anticipate an actor's actions when these actions can be predicted only by attributing a false belief to the actor" (p. 587).

It has been known for some time that very young children understand others' goals, desires or intentions (Wellman, 1990; Woodward, Sommerville, & Guajardo, 2001). This tallies with interesting findings in first-language acquisition. Cross-linguistically, children use predicates expressing volative modality, such as 'want', before words expressing alethic modality, such as 'believe' (Tsimpli & Smith, 1998, p. 197).

Commenting on the work of Woodward *et al*, Malle, Moses and Baldwin write that "infants as young as 9 months understand the goal-oriented quality of some intentional actions" (Malle, Moses, & Baldwin, 2001, p. 11) and that:

> at this early age infants already use information external to the behaviour stream to determine the relevance of a goal object. For example, previously provided information about an agent's interest in the contents of a box led infants to construe a subsequent box-grasping action as goal-oriented; in the absence of such prior information, infants failed to register the action's goal-oriented quality. (Malle, Moses, & Baldwin, 2001, p. 11)

Corroboration that nine-month-olds grasp intentions comes in work from Behne and colleagues who found that children from this age up were more impatient with an adult's unwillingness to perform an action than with failure through inability (Behne, Carpenter, Call, & Tomasello, 2005).

A great deal of work has been done on imitation. Children's imitation of others' actions is apparently aimed at reproducing the physical behaviour from birth, but it is reoriented to the goal of that behaviour from around 18 months. For example, if the experimenter has his hands full and uses his head to turn on a light, the child turns it on with his hand, imitating the goal rather than the means used thus indicating an understanding of another's intentions (Meltzoff, 1988; Meltzoff & Brooks, 2001, p. 13).

Returning to children's knowledge of others' beliefs, Onishi and Baillargeon (2005) found that children as young as 15 months seem to keep track of where an agent thinks an object is, reacting with surprise when an adult looks in a place to which the child, but not the agent, has seen the object moved. These results suggest that children as young as it is possible to test are keeping track of some mental states of others.

While Onishi and Baillargeon note a possible alternative explanation – that children keep track of what others have *seen*, rather than what they believe, and expect them to look for an object where it was last seen – they prefer the explanation that children keep track of others' beliefs. They have some evidence for this view:

> Recent results of ours have indicated that infants can predict where an actor will search for a hidden toy even when she does not see it disappear but must infer its location based on various (useful or misleading) cues. (Onishi & Baillargeon, 2005, p. 257)

Surian, Caldi and Sperber (2007) found that 13-month-olds watching agents search for objects were surprised when search was effective if and only if the agent had not had access to relevant information. As they say, this "supports the view that infants possess an incipient metarepresentational ability that permits them to attribute beliefs to agents." (p. 580)

One can imagine the two types of explanation for such results coming apart in circumstances where it is obvious that an agent would not expect to find the object in the place where she last saw it. Would a child be surprised if an agent did not look for a cork where she left it in a stream, for example? I suspect not.

Here, too, a determined opponent of a representational theory might be able to resist the conclusion that children keep track of others' beliefs, but only by postulating increasingly complex (and *ad hoc*) rules of thumb that govern children's expectations, so that they act as though they kept track of beliefs without actually doing so.

The best current alternative for those who resist crediting infants with the ability to act on others' knowledge appears to be the theory that they understand that behaviour is aimed at goals, and they expect agents to attempt to achieve those goals in a direct way, that is, "Teleological understanding, in which behavior is understood as being due to goals and external circumstance (true beliefs), and a rationality assumption is made that the most efficient means of achieving the goal are taken." (Ruffman & Perner, 2005, p. 462). Gergely and colleagues (2002) reinterpret Meltzoff's findings about imitation in this way. There is recent evidence that does not fit easily with this theory, however, from work on communication.

It has been known for some time that communicative tasks can be a facilitating factor for behaviour that takes into account others' mental states. It is true that young children give the wrong referent on a modified false-belief task with referring expressions (the 'message-desire discrepant task') (Mitchell, Robinson, & Thompson, 1999), but very young children have been known for some time to modify their communicative behaviour to take account of the knowledge of others (O'Neill, 1996).

Tomasello and colleagues argue that humans, unlike the other apes, share, and that this extends to communication: we share information as naturally as we share objects such as food (Tomasello, Carpenter, Call, Behne, & Moll, 2005; see also Enfield & Levinson, 2006, p. 26). Sharing of information starts young and takes into account the knowledge and goals of the hearer (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998).

In a series of experiments (Liszkowski, Carpenter, Henning, Striano, & Tomasello, 2004; Liszkowski, 2005; Liszkowski, Carpenter, Striano, & Tomasello, 2006; Liszkowski, 2006; Liszkowski, Carpenter, & Tomasello, 2007), Liszkowski and collaborators have shown that year-old children point to establish shared attention to a referent, that is, to communicate. What is more, if the experimenter appears to misunderstand, the child will point again, repeatedly. As Enfield and Levinson comment:

> This is a spectacular finding, because TOM literature standardly suggests that the ability crucial to this account (i.e., knowing that the other does not know something) is a much later achievement in development, coming not at 12 months but at four years. In Liszkowski's studies, the child is clearly using pointing for informing, one of the main motivations for communication. (Enfield & Levinson, 2006, p. 16)

Not all theorists agree that the results to date demonstrate that infants share information (see Tomasello, Carpenter, & Liszkowski, 2007 for a recent defence of this interpretation), but there is some consensus, even among some who oppose this view, that year-old children do engage in genuinely communicative acts intended to direct attention (Southgate, van Maanen, & Csibra, 2007). If the general tendency of these results is borne out, one motivation for non-intentional theories of communication will be weakened or removed.

In their response to Millikan's comment about mindreading, Sperber and Wilson (1987a) acknowledged that whether young children have mindreading abilities is open to further investigation. I think, to summarize, that while that is still true twenty years on, there is increasing evidence that young children do keep track of others' mental states and take account of them in their actions in various ways. Their abilities do not extend to passing the rather elaborate verbal false-belief task, nor to "conscious metacognitive inferences" or the ability to "articulate a conception of beliefs as truth-evaluable mental states" (Surian, Caldi, & Sperber, 2007, p. 585), but they do involve, even at a very young age, tailoring utterances so that they are suitable for a hearer given what the child knows about the speaker's goals and what the speaker has seen.

Even sceptical theorists now accept infants have rather complex abilities, including a working assumption that goals will be achieved efficiently, and a means of keeping track of what others have seen. If in fact communicative ability in infants makes use of abilities of this sort, even to the exclusion of explicit representations of others' beliefs, then it can be seen as inferential, as long as we grant that simple heuristics, such as working on the assumption that an agent knows what she has seen, can play a role in inference processes.

Pragmatic development

There is a further question for anti-intentionalists. The assumption is that young children do not have mindreading ability, but can communicate. Therefore theorists postulate communication ability that does not require theory of mind. This form of communication would be rather basic, so infants might be able to perform reference assignment and disambiguation, but not to work out implicatures.

However, children's pragmatic abilities fall well short of adults' until much later than four years old, the point at which they pass the false-belief task. Adult levels of performance with metaphor and idiom, for example, come much later. Mitchell and colleagues write that, "Considerable development in the ability to distinguish between literal and intended meaning seems to occur around the age of 6–8 years" (Mitchell, Robinson, & Thompson, 1999, citing a number of studies). According to Winner and colleagues, the ability to choose a metaphorical versus a literal interpretation increases from 6 to 9 years old and again from 9 to 14. (Winner, Engel, & Gardner, 1980; Winner, Rosentiel, & Gardner, 1976; Winner, 1988)

A second example is ability with so-called 'scalar' implicatures. Utterances of sentences such as the one in (5a) can convey a meaning like the one in (5b):

(5) a) Some of the linguists danced.

b) *Some and not all of the linguists danced*

Noveck (2001) has shown that on tasks probing this ability children answer semantically rather than pragmatically: they would take (5a) to mean that *some (and possibly all) of the linguists danced.* They are still below adult performance at ten years old.

Therefore explanations are needed for lower-than-adult performance on pragmatic tasks for children up to ten or even fourteen years old. Such explanations might also be capable of explaining pragmatic deficits in children younger than four. So it is not clear that it is necessary to invoke lack of mindreading abilities to explain pragmatic deficits in young children.

Another way of putting this point is to say that the developmental pragmatics literature as a whole does not support the theory that there is a radical discontinuity in children's pragmatic abilities around four years old. Instead there is a long, slow increase in children's pragmatic performance, with, on the one hand, the ability to tailor utterances to hearers in infants as young as can be tested and, on the other, some studies showing difficulties with figurative speech as late as the mid-teens.

Pragmatic processing depends on world knowledge (by definition), on processing and attentional capacity, and on strategies for dealing with information. Children certainly develop all of these from infancy, continuing into their teens. Increasing processing capacity and incremental adjustment of memory, together with an innate ability to represent speakers' meanings and other mental states seems as promising an explanation as any for the data. Of course, the gradually accumulating knowledge and strategies that this picture suggests need not be conscious or explicitly mentally represented. Much of the increase in pragmatic ability could be due to adjustment of accessibilities of concepts from lexical items and other pieces of information. I comment further in chapters 3 and 5 on the role of this kind of attunement to the problem domain in heuristic searches.

1.4 SUMMARY

In this introductory chapter I have sketched out the thesis that I intend to defend and some of the assumptions that I make in doing so. The thesis concerns the role of rationality in utterance interpretation and utterance production. I accept the broad outlines of Grice's picture of language use as a rational activity. An utterance brings about an interpretation by triggering a reasoning process in the mind of the hearer. The process seeks an explanation for the utterance (and the way it is made in a particular context) in terms of the speaker's intentions, reaching a representation similar to the one in example (3) above.

I have already raised the question of the status of explanations like this. For a psychological theory to be explanatory it must deal with the causes of behaviour, as Millikan says. That is, it must give an explanation in terms of mental processes and mental states.

Realistic theories of human cognition propose that much of it involves fast and frugal shortcuts. This must also be true of pragmatic processing. I agree with Sperber and Wilson, who wrote two decades ago that:

> ... if there is one conclusion to be drawn from work on artificial intelligence, it is that most cognitive processes are so complex that they must be modelled in terms of heuristics rather than failsafe algorithms. We as-

> sume, then, that communication is governed by a less-than perfect heuristic. (Sperber & Wilson, 1986, p. 44)

Gricean inference schemas resemble logical arguments, not heuristic shortcuts, however. Is there a problem here? It has certainly seemed so to many theorists. One way of seeing this is as a clash between the demands of ecological rationality – how well suited a procedure is to a particular type of problem – and more traditional notions of rationality. In this introduction I have set aside alternatives that would lead away from this knot, particularly non-inferential theories of utterance interpretation, but also non-mentalist theories of thought.

In the next chapter, I outline a traditional view of rationality and reasoning according to which reasoning ability is the ability to make reason-preserving transitions. This might be thought to deepen the explanatory gap, but I argue that a theory of this kind is compatible with the view that much reasoning is carried out a 'quick way' (as Grice puts it). In chapter 3 I look in some detail at the reasons for theories of rationality as bounded, then show that heuristic search is a good candidate for the quick way of reasoning in the case of inference to the best explanation.

I return to inferential-intentional theories of communication in chapter 4 and the details of my view of reasoning in pragmatic processing (primarily utterance interpretation) emerge there and in the final chapter.

My view of how broadly Gricean pragmatics is realised in the mind is essentially a version or interpretation of Sperber and Wilson's relevance-theoretic comprehension procedure. What is new here is the attempt to show that on current views of pragmatics and of reasoning, options are limited, and a heuristic with many of the properties of the relevance-theoretic comprehension procedure is a natural conclusion.

It is worth noting that I hold a somewhat different view to Sperber (Sperber, 2000; Sperber, 2001) on what reasoning is. In my opinion it is not necessarily a metalevel process, but any process that takes conceptual input and aims at the preservation of rational value. On my view reasoning (usually fast) is the ordinary business of many central processes, rather than something reserved for a module that represents the output of such processes.

Chapter 2 · Rationality and inference

> when most of us talk of reasoning, we think of an occasional, conscious, difficult, and rather slow mental activity. What modern psychology has shown is that something like reasoning goes on all the time – unconsciously, painlessly, and fast (Sperber, 1995, p. 195)

2.1 INTRODUCTION

In chapter 1, I discussed the view of pragmatics that I want to establish, setting out some initial reasons for seeing utterance comprehension as thoroughly rational yet performed by heuristics. In chapters two and three, I look in more depth at two competing visions of rationality, classical and bounded rationality, aiming to develop a view of what kind of rationality should be attributed to people and how it might be understood scientifically. The aim is not a full definition of rationality, a notoriously slippery and fundamental concept, but enough of a characterisation of the area to work with when discussing utterance understanding as a rational activity.

This project has a good deal in common with Grice's investigation of rationality (Grice, 2001). The main goal of the next two chapters is to integrate on the one hand Grice's views that a necessary condition for rationality is a certain minimum reasoning ability, and that reasoning should be characterised as an activity that aims at value-preserving transitions between inputs and outputs, with, on the other, views of rationality and reasoning as bounded by human cognitive limitations of time, effort and working memory capacity, and other limitations on mental representations or mental processing. (Such views are held by Simon (1957b), Cherniak (1986), Sperber and Wilson (1986), and Gigerenzer and colleagues (e.g. Gigerenzer & Todd, 1999)). This is more a matter of bringing out certain possibilities in Grice's account than of disagreeing with it, since Grice allows that reasoning is generally not spelled out laboriously step by step, as mentioned in the introduction and discussed further here. Both of these views are necessary for a realistic inferential-intentional view of pragmatics as outlined in the introduction and discussed fully in chapter 4.

Human reasoning has received a great deal of attention from psychologists, particularly since the move to cognitive psychology in the second half of the twentieth century. Two long-standing debates in the psychology of reasoning are relevant. They are briefly introduced here and discussed further in the body of this chapter. One is the debate about whether systematic and reproducible errors on reasoning tasks put in doubt the traditional idea that humans are rational. The arguments have been heated, but it seems that a consensus is emerging that humans are neither systematically irrational nor as normatively rational as some once assumed (Samuels, Stich, & Bishop, 2002; Samuels & Stich, 2004). The conclusions we reach are not always warranted and we do better or worse depending on the form of the task and how information is presented. In many cases, what is striking is how well we manage in a short time with limited information and mental resources.

The second relevant debate in the psychology of reasoning is the controversy about whether deductive reasoning is carried out by following rules of derivation akin to those used in logical derivations (Braine, 1978; Braine, Reiser, & Rumain, 1984; Braine & O'Brien, 1998; Rips, 1983; Rips, 1994; Rips, 1997), or whether the method used is the construction of mental models of states of affairs, which yield conclusions on examination (Johnson-Laird, 1983; Johnson-Laird, 1999; Johnson-Laird, Girotto, & Legrenzi, 2003) – or neither of these. I tentatively follow Sperber and Wilson's endorsement (Sperber & Wilson, 1986, pp. 102–103) of a mixed picture, with some deductive, truth-preserving inference rules sensitive only to logical form, some rules based on conceptual information (meaning postulates) and some additional inferential procedures, perhaps including mental models.

This debate has mainly focussed on deductive reasoning with a closed set of premises and on performance on certain reasoning tasks (this is stated explicitly as the aim in Braine, 1978; and noted by Sperber & Wilson, 1986, p. 97). However, it is plausible that the resources available for deduction are put to use in the distinct task of generating inferences from new (or newly presented) information (Sperber & Wilson, 1986, p. 97). If further information from working memory or the environment can be introduced as additional premises, then non-demonstrative inference can be modelled with no need for special non-demonstrative 'inference rules' (Sperber & Wilson, 1986, pp. 107–117).

I reserve for chapter 4 discussion of a third debate about reasoning which has received a great deal of attention in recent years: whether explicit, conscious reasoning and unconscious reasoning rely on qualitatively different mental processing.

Thus far, I have assumed that accounts of human reasoning ability can and should rely on a realistic view of mental representations. That is, it is assumed that 1) the mind is (among other things) an information-processing system in which information is mentally represented, and that 2) the form in which a piece of information is represented in the mind has a strong effect on what can be done with it and what is likely to be done with it. The form of a mental representation determines what other pieces of information it can interact with[28] and therefore what further information can be derived from it. That is, something like[29] Fodor's Representational Theory of Mind (RTM) (Fodor, 1975) is presupposed. I examine the reasons for working with a view of reasoning of this kind.

It is a well-known irony that Fodor's theory has been applied to central processes and non-demonstrative reasoning, areas in which Fodor believes no progress can be made in this (or any other current) framework (Fodor, 1983, p. 107). Fodor justifies his scepticism by pointing out that central belief formation is sensitive to 'global factors'. No one knows, he claims, how such factors affect the process. (Fodor, 1983, p. 129) (See Fodor, 2000; Fodor, 2005 for Fodor's continuing scepticism about the cognitive science of central processes.) I will call the question of how central cognition, particularly abductive

[28] What other pieces of information it actually does interact with is presumably partly determined by the accessibility and activation of those other pieces of information, and partly by what other processing is competing for mental resources.

[29] I write 'something like' Fodor's theory because his theory also comes with a particular view of the semantics of concepts in the language of thought. I avoid this issue.

inference, can be modelled computationally *Fodor's problem* (following Carruthers, 2003).[30]

I believe that such deep pessimism is unwarranted. In section 3 of this chapter, I discuss RTM and suggest that one reason for Fodor's scepticism is his exclusive stress on the propositional (or logical) in reasoning. This neglects another important stream of research on reasoning particularly stressed by Simon: reasoning as problem-solving (Simon, 1990, pp. 11–13), which models reasoning as sequential generation and assessment of trial solutions. Both are indispensable, I argue, for some aspects of central cognition, including those involved in abductive inference, such as the inferential aspects of utterance understanding.

The short timescale of utterance interpretation and the fact that communicative inputs generally come from a helpful source (Sperber & Wilson, 1986, pp. 66–67), and the tight fit between the structure of the environment and the heuristic applied are further considerations bearing on Fodor's argument as it relates to pragmatics. I return to these points in chapter 5 in which I put into practice some of the consequences of the discussion of rationality in the next two chapters, considering the degree to which pragmatic processing can be both modular and central.

In the next section of this chapter, however, I put aside the issues of utterance interpretation and of mental representation, focussing on a traditional view of rationality endorsed and refined by Grice.

2.2 RATIONALITY AND REASONING

> if, as it seems not unreasonable to suppose, reason is, as of its nature, the faculty which is manifested in reasoning, then it would be a good idea to investigate what reasoning is. (Grice, 2001, p. 5)

I adopt Grice's suppositions (1) that rationality is the possession of reasoning ability and (2) that reasoning is an activity aimed at making value-preserving transitions, so that reasoning leads from premises to conclusion like a logical

[30] I avoid the name Fodor uses, the 'frame problem', because that is arguably the name of a different problem (Hayes, 1987).

argument. Among philosophers, the second of these views, although traditional, is controversial. More neutral characterisations are often given. Gilbert Harman, for example, has advocated a broader picture of reasoning as 'change in view', emphasising the conceptual distinction between laws of inference (logical) and rules or procedures for reasoning (psychological) (Harman, 1984; Harman, 1986). Reasoning, according to this account, is much *more* than stringing together truth-preserving transitions. I follow Grice in attempting to set aside this kind of objection by considering inferential ability as the core of reasoning, albeit not the whole story (see section 2.2.3 below).

A more pressing concern, in my opinion, is that a traditional view of reasoning may not translate well into a realistic theory of cognition, given that what actually happens in reasoning must often make use of heuristic shortcuts rather than truth-preserving rules. According to this objection, reasoning is often *less* than stringing together truth-preserving transitions. I think that the evidence is indeed compelling that heuristics play a key role in reasoning, and that – therefore – the treatment of reasoning in cognitive science needs to take account of this. I argue that Grice's picture of reasoning, although perhaps agnostic about mental representation, provides a way of answering this criticism. The idea is that some episodes that skip many of the required truth-preserving steps are nonetheless reasoning, due to the intended resemblance of the activity to the construction or rehearsal of an argument. (See section 2.2.5 below.)

A central element of Grice's picture of reasoning, which I comment on but do not commit myself to, is connected to the traditional distinction between theoretical and practical reasoning. Grice wanted to make plausible the idea that there is a close parallel between reasoning about what is true and reasoning about what is to be done, that is, between theoretical and practical reasoning (and perhaps also other types of reasoning, if there are such). Grice sketches out a unitary account according to which reasoning in all domains is value-preserving, where the value preserved may be different for each domain: truth in the theoretical domain and practical goodness in the practical domain. In this thesis I am mainly concerned with theoretical reasoning, but below I outline the distinction between theoretical and practical reasoning and comment briefly on its relevance to pragmatics, and on Grice's attempted unification.

The value that theoretical reasoning is usually seen as attempting to preserve is truth, since that is the value possessed by propositions and preserved by deductive inferences. There are two ways in which this assumption might need to be relaxed. Sperber and Wilson make a convincing case that in cognitive science it is necessary to allow for inferences operating over representations that are syntactically well-formed but semantically incomplete in the sense that they fall short of propositionality (Sperber & Wilson, 1986, p. 72f). These inferences can be made with the same rules that are truth-preserving when applied to fully propositional thoughts.

Not every belief is held with certainty, and a second relaxation to the model of reasoning may need to be made to accommodate this fact. Reasoning from two or more beliefs that are less than certain generally yields a conclusion that is also less than certain. Inferences from uncertain knowledge can still be treated as value-preserving where the value preserved is *warrant* (Sperber & Wilson, 1986, p. 108ff). A belief is supported to a certain degree by the beliefs it is deduced from. How much support is provided depends on the certainty of the premises. The degree of certainty can be seen as ranging from one (certain) to zero (certainly false). Then, logically, the support provided for a deduced conclusion is the product of the warrant of each premise[31]. A conclusion is at least as warranted as the support it receives from its premises indicates – perhaps more, because there may be other evidence in its favour, but not less.

These facts about warrant are no bar to a picture of reasoning as ability with truth-preserving inferential rules. Inferential rules that meet Grice's criterion that they preserve truth if the inputs they operate on are true will preserve warrant when used with beliefs that are less than certain.

[31] I am not suggesting that we assign a numerical probability to each belief that we hold. That is implausible *a priori* and not supported by experimental evidence. I agree with Sperber and Wilson that we are able to make non-numerical estimations of the degree of certainty we assign to propositions (such as *certain, highly probable, possible, unlikely,* and *certainly false)*; and that estimates of probability are not generally comparable across domains (Sperber & Wilson, 1986, pp. 77–81).

> Let us, then, take as a first approximation to an account of reasoning the following: reasoning consists of the entertainment (and often acceptance) in thought or in speech of a set of initial ideas (propositions), together with a sequence of ideas each of which is derivable by an acceptable principle of inference from its predecessors in the set. (Grice, 2001, p. 5)

The idea that reasoning involves making steps that preserve truth, as in a valid logical argument, is, as remarked above, quite traditional. A recent paper by Michael Smith (2004) attributes a view of this kind to Hume[32]. Thomas Reid had similar views on this subject:

> In all reasoning ... there must be a proposition inferred, and one or more from which it is inferred. And this power of inferring, or drawing a conclusion, is only another name for reasoning: the proposition inferred being called the *conclusion*, and the proposition or propositions from which it is inferred, the *premises*.
>
> Reasoning may consist of many steps; the first conclusion being a premise to the second, that to a third, and so on till we come to the last conclusion." (Reid, 1855, p. 424)

Harold Brown (1988) outlines a related "classical theory of rationality" (as a contrasting background to his own views) whose essential feature is that reasoning makes use of algorithms, procedures that are guaranteed to arrive at the right answer, given the right input.

The theory is that, in correct reasoning, the beliefs that a reasoner starts with logically support the belief or beliefs he reaches, in just the same way as the premises support the conclusions of a valid argument.

> Suppose that I begin by believing that *p*, and believing that if *p* then *q*, and on the basis of these beliefs, come rationally to believe that *q*. The obvious

[32] Fodor might disagree with this attribution. He attributes to Hume the view that mental states cause other mental states *through laws of association* (Fodor, 1983, pp. 27–8, 31). On this view, Hume's theory was realist about mental causation but lacked the technology (laws of natural deduction) needed to make such a theory work. See section 2.4 below for more on this.

> explanation of the rational transition between my beliefs, is that, *inter alia*, there is an isomorphism between their relations and the logical relations between the propositions I believe, that is, the propositions that give the reasons why *q* ... (Smith, 2004, p. 77)

On this view, reasoning essentially involves constructing – or rehearsing – sequences of states that parallel valid arguments. A simple example is given in table 1.

Table 1 (Smith's 5.1): Parallel between psychological and logical inference

Of course, reasoning can be much more complex than simple application of *modus ponens*. According to the present view of reasoning, greater complexity in reasoning is primarily due to the joining together of simple steps, in just the same way that a complex logical argument can be built up from repeated applications of the rules of natural deduction. Since the rules of natural deduction preserve truth (by definition) the output of the reasoning process will be true if the input was true. The reasoner will not go wrong in believing the output proposition, then, if he was not wrong in believing the input propositions.

More generally, as noted above, one could try to extend this picture into other domains of reasoning by the postulate that in all domains the transitions preserve value of some kind: truth in the theoretical domain, and other kinds of value in other domains. I return to this point in the discussion of theoretical and practical rationality below.

I also noted above that some generalisations of this sort may be necessary even in the theoretical domain, since it seems that what is entertained in reasoning may include thoughts that fall short of being propositional and beliefs that are less than certain. Thoughts which are not fully propositional, while well-formed, would be semantically incomplete in the sense of lacking truth-conditions and thus truth-values. It cannot be, then, that truth is preserved in reasoning from such thoughts. However, as suggested above, there is no contradiction with the spirit of Grice's account since no special rules or procedures are needed, just standard inference rules that are truth-preserving when given fully propositional input. I reserve further comment on this issue to section 2.4 below, since it is easier to discuss the need for this generalisation in the context of a realistic theory of mental representation[33], which the present discussion does not presuppose.

Reasoning with beliefs that are less than certain can fall under the value-preserving generalisation as long as there is some kind of value preserved by valid inferences from both certain and uncertain beliefs. As noted above, this value is *warrant.* If a rule is truth-preserving then it is also, *ipso facto,* warrant-preserving. An inference using truth-preserving rules from a set of beliefs held with varying degrees of certainty provides some support for a conclusion or conclusions, in proportion to how certain each initial belief is. In the special case when all of the initial beliefs are certain then the conclusion or conclusions are also certain.

Just as it is traditional to see the rationality of human beings as centrally involving the possession of reasoning ability, it is also traditional to see it as centrally involving the ability to work with reasons. As Grice remarks, "the connection between the two ideas is not accidental" if one accepts the present view of reasoning as entertaining or generating chains of thoughts linked by value-preserving transitions (Grice, 2001, p. 5). According to the theory that reasoning involves only steps that preserve acceptability, if a reasoner starts off with reasons for accepting the initial set of thoughts, then he has reasons for accepting the conclusions which are derived from those thoughts.

It is worth commenting briefly on the history of the view of reasoning I have been setting out. It is hard to overstate how traditional this view is. Both Boole and Mill, in their classic works on logic, aimed to contribute to under-

[33] That is, a theory that claims that mental representations are real and that they are perfectly respectable entities to appeal to in scientific accounts, a view mentioned in chapter 1. I discuss and endorse this theory as applied to conceptual representations in section 2.3 below. I am not committed to another sort of realism about mental representations which claims that conceptual and/or perceptual mental representations have intentional properties. In other words, I use the term 'representation' in Chomsky's broad sense.

standing of the laws of thought, where 'laws of thought' is understood in the strong sense of the laws that thought follows, rather than normative laws of logic that attempts at reasoning can be measured against. Boole set himself the task (in the first paragraph – and in the title – of his 'Investigation of the Laws of Thought'), "to investigate the fundamental laws of those operations of the mind by which reasoning is performed" (1854), and to draw more general conclusions about the mind if possible: "to collect from the various elements of truth brought to view in the course of these inquiries some probable intimations concerning the nature and constitution of the human mind." Mill's objectives were similar: "Our object, then, will be, to attempt a correct analysis of the Intellectual Process called Reasoning or Inference, and of such other mental operations as are intended to facilitate this..." (1856, p. 7) Both works aimed at developing formal systems, to be sure, but this was to be accomplished by investigation of the way we actually think[34] [35].

Against this background, Frege made a clear distinction between logic and psychology. Logic is the study of the laws of truth, whereas psychology is the study of the laws of thought, including reasoning. Logic does not depend on psychology, but psychology has to heed logic, since logical laws are normative for reasoning, given that reasoning aims at truth: "Like ethics, logic can also be called a normative science. How must I think in order to reach the goal, truth? We expect logic to give us the answer to this question" (Frege, 1979, p. 128).

The comparison with ethics needs to be put in context. It seems that Frege did not ultimately think that logic was normative in the same way as ethical laws:

> The word 'law' is used in two senses. When we speak of moral or civil laws we mean prescriptions, which ought to be obeyed but with which actual occurrences are not always in conformity. Laws of nature are general fea-

[34] The roots of this approach are to be found in "the common eighteenth-century equation between logic and grammar" (Wallace, 1980, p. 341), itself with roots in the perfect language tradition (Walker, 1972; Land, 1974). Coleridge's early nineteenth century work on logic and the philosophy of language (for which see Wallace, op cit.) looks back in this direction and forward to Boole and Mill.

[35] Gigerenzer and Hoffrager (1995) provide a brief survey of the related Enlightenment conception of laws of *probability* as laws of the mind. See chapter 3 for more on this.

> tures of what happens in nature, and occurrences in nature are always in accordance with them. It is rather in this sense that I speak of laws of truth. Here of course it is not a matter of what happens but of what is. From the laws of truth there follow prescriptions about asserting, thinking, judging, inferring. (Frege, 1984, p. 351)

According to this view, logic, like physics, mathematics and psychology, is normative in its own field: each of these subjects tells us how certain kinds of things are, and therefore how we ought to think about those kinds of things. Macbeth summarises Frege's "considered view":

> Any science that aims to discover laws rather than facts (for example, the facts of natural history) is normative in a sense: insofar as it discovers laws governing what is, it also sets out prescriptions governing our thoughts, judgments, and inferences regarding what is. (Macbeth, 2005, p. 23)

The rules of logic are more general than the rules of (e.g.) physics, though. Whereas the laws of physics tell us how we must think about physics, the laws of logic "are the most general laws, which prescribe universally the way in which one ought to think if one is to think at all". (Frege, 1964, pp. 12-13).

There may have been a tension in Frege's thought between two ways of seeing the laws of logic. One is to see them as the laws of truth: they tell you how to think if you want to think true thoughts. The other is as the laws of truth-preservation: they tell us which inferences preserve truth, and thus they tell you how to think if you want your conclusions to follow from your premises. Since this thesis is not concerned to define logic, I adopt the latter view with no further comment.

Granting Frege's point that facts about psychology do not determine facts about logic, it is tempting to wonder whether the general acceptance of this point had a damping effect on the study of the psychology of reasoning, at least within philosophy. Once it was unfashionable to see logic as the grammar of thought, attempting to discover the laws of logic by investigating how people reason was less attractive. Braine, following Henle (1962) in this respect, claims that "this change of stance reflected a changed intellectual climate, not any fresh insight into the nature of reasoning" (Braine, 1978, p. 2). Further examining this claim about the history of philosophy would require too lengthy a digression for this thesis. In any case, the debate has since partly shifted into the psychology of reasoning and become partly empirical – or one might say that logic, the psychology of reasoning and philosophical treatments of rationality have become established as three separate fields. Within the psychology of reasoning, the mental logic programme can be seen as the revival of aspects of the traditional view. This programme presupposes a realist view of mental representation (which I share: see footnote 33 above). I return to this issue in section 2.3 below. In the current section I continue to explore the Gricean version of the traditional theory.

It is perhaps a consequence of the traditional picture of reasoning that the words we use to speak about reasoning and about logical inference are not clearly distinguished. We find it at least as natural to apply the words 'deduction' and 'inference' to instances of reasoning as to derivations of logical sequents. Logicians call the rules employed in syntactic derivations of logical sequents the 'laws of natural deduction', but that is also a good name for the psychological rules of a mental logic. In ordinary speech we use 'conclusions' to refer to the propositions derived from a process of reasoning as readily as to refer to propositions entailed by some premises (and the verb 'conclude' also works in both contexts)[36]. From these informal observations about meaning and usage, of course, nothing follows for the truth of the traditional picture, but they are at least indicative of its familiarity.

2.2.2 NON-MONOTONICITY, ABDUCTION AND INDUCTION

There is room for doubt about whether this picture of reasoning has much generality. Notoriously, only deductive reasoning could be purely a matter of making truth-preserving transitions. Deductive reasoning can be defined as reasoning that aims to work out what necessarily follows from a closed set of propositions. Given the parallel with deductive logic, which is the study of logical necessity, it is not entirely surprising that deductive reasoning seems to fit the traditional picture. However, even within deductive reasoning there are cases where the parallel is not so clearly preserved.

[36] The word 'premise', on the other hand, is more at home in logic than in talk about reasoning. This is presumably related to the fact that it is a more technical word than 'conclusion'.

One reason to doubt the neat connection between deductive logic and deductive reasoning is what is sometimes called the non-monotonic character of reasoning. Work on defeasible logics and non-monotonic reasoning is motivated by the observation that the conclusions of some ordinary deductive inferences are typically withdrawn when new information is presented. One is told that "If the switch is down, the light is on" and "The switch is down" and one concludes that the light is on, but would withdraw this conclusion if told that there is a power cut. One might (but might not) withdraw the conclusion if told that it is true that "If there is not a power cut then the light is on".[37] The initial reasoning parallels the logical rule of *modus ponens*, but the revision of the conclusion is not so easily explained in these terms, since it is a property of logical inferences (in classical logic) that adding an extra premise to the set of premises does not (that is, cannot) remove any conclusions from the set of conclusions. This property can be called monotonicity. Defeasible logics, and their instantiation in research in computer science on non-monotonic reasoning (e.g. Antoniou & Williams, 1997), are formal solutions that aim to preserve the parallel between logic and reasoning by doing without the property of monotonicity.

Despite the formal work, however, there is no settled theory of how people revise conclusions in the light of extra information. As the psychologist of reasoning Johnson-Laird says, "Philosophers and artificial intelligencers formulate such systems of 'defeasible' or 'nonmonotonic' reasoning but psychologists do not know how people reason in this way." (Johnson-Laird, 1999, p. 112). One obvious avenue to explore is that deductive inferences are made in accordance with classical, monotonic logic and that when conclusions of such deductions are withdrawn in the light of new information this is because the new information casts doubt on the premises. This view, which I support, contrasts with the view of reasoning as non-monotonic. According to the non-monotonic–reasoning view, the withdrawal of the conclusion is because the new information, when added to the original premises, undermines the inference itself. I think that consideration of examples suggests that new in-

[37] In the psychology of reasoning this phenomenon is known as the Suppression Effect (Byrne, 1989; Byrne, 1991), particularly when the conclusion is withdrawn in the light of an extra conditional.

formation undermines belief in the original premise(s) rather than in the inference. In the example given above, the conclusion 'the light is on' is withdrawn because the new information 'If there is not a power cut then the light is on' causes the reasoner to doubt the original premise 'If the switch is down, the light is on'.[38] Byrne, Espino and Santamaría (1999; 2000) and Politzer and Bourmaud (2002) have given related explanations (although slightly different from each other) of the withdrawal of the conclusion of a deductive inference.

These approaches share a presumption with work on non-monotonic reasoning that logic and reasoning should be kept in step with each other. A different way of proceeding is to deny the traditional view that reasoning is primarily a matter of truth-preserving steps. Advocates of this alternative can also point to more glaring differences between logic and non-deductive reasoning.

The most striking disanalogy of this type concerns non-demonstrative inference. Abductive reasoning and inductive reasoning both aim at reaching conclusions that are not logically entailed by the starting points taken as premises, that is, they both involve non-demonstrative inference. Abductive reasoning is inference to the best explanation of some observation or fact; inductive inference makes a generalisation from several observations or facts to a covering law or regularity.[39]

The close parallel between a logical argument and the reasoning process seems to break down for these forms of reasoning. In non-demonstrative inference, by definition, the propositions that one starts out believing do not entail the proposition that one ends up believing. As Smith puts it: "the hallmark of inductive reasons – reasons such as those provided by the consideration that something or other is the best explanation of some aspect of our experience – is precisely that they do not logically entail the conclusions that we think they are reasons for"[40] (Smith, 2004, p. 79). For example the proposition

[38] The mechanism might be *reductio ad absurdum*, particularly in the case when one is told that there is a power cut. In the context, being told that there is a power cut implicates that if there is a power cut the light is not on, since the information would be irrelevant and misleading otherwise. Then a contradiction can be derived: the light is on and the light is not on; so the first conditional (if the switch is down the light is on) is discarded, and along with it, the proposition that the light is on.

[39] Abduction is sometimes regarded as a species of induction.

[40] As the quotation indicates, Smith does not distinguish in the cited paper between induc-

that *the barometer is falling* does not entail the proposition that *it will rain tomorrow.* Nor is the proposition that *it will rain tomorrow* entailed by the propositions that *the barometer is falling* and that *the best explanation for the barometer falling is that something is happening that means it will rain tomorrow* (Smith, 2004, p. 79).

Thus there is an obvious disanalogy between the laws of natural deduction in logic and the steps taken in inductive and abductive reasoning. In order to preserve the parallel in the domain of non-demonstrative reasoning it might seem that there would have to be transitions in this area which meet Grice's criteria. That is, what we are looking for are:

> forms of transition, from a set of acceptances to a further acceptance, which are such as to ensure the transmission of value from premisses to conclusion, should such value attach to the premisses. (Grice, 2001, pp. 87–88)

In other words, there would have to be laws of non-demonstrative inference whose application to some input yields output that preserves the rational acceptability possessed by the input. There are no such transitions that are generally accepted: "There is no well-developed system of inductive logic that would provide us with a plausible model of the central cognitive processes." (Sperber & Wilson, 1986, p. 67)

It is often said that this contrast with deduction is the explanation of Hume's well-known scepticism about inductive reasoning (Cohen, 1992; Smith, 2004)[41]. Cohen writes:

> Hume assumed the only valid standards of cognitive rationality were ... deductive, mathematical or semantical[42]. Induction was not a rational procedure, on his view, because it could not be reduced to the exercise of reason in one or another of these three roles. (1992, p. 417)

Given a picture of reasoning as essentially involving transitions that preserve truth, it is certainly harder to accommodate abduction or induction than de-

tion and abduction.

[41] Smith also explains Hume's scepticism about practical reasoning in these terms.

[42] By 'semantical', Cohen means inferences that depend on non-logical lexical items: from *Teddy is a cat* to *Teddy is a mammal*, for example.

duction. Thus, one might think that if we accept such a picture we should be sceptical about the viability of non-demonstrative reasoning, as Hume famously was.

Similar doubts have been raised by critics of Grice. In a review of Grice, 2001, Harman writes:

> Reasoning may sometimes involve constructing an argument, but not always because one is reasoning from the premises of that argument. The argument is often an explanatory argument, and one is reasoning from the conclusion of that explanatory argument to a conclusion that is a premise of the argument. (Harman, 2003)[43]

I think Grice's picture of reasoning and rationality is worth defending against this kind of objection. In fact, since I also accept Grice's characterisation of the fundamentals of communication, I need to show how some abductive reasoning at least is compatible with the traditional picture of reasoning. As I explained in chapter 1, in a broadly Gricean account of communication, hearers reason from (facts about) utterances to speaker's intentions, where the intentions are explanations for (the facts about) the utterance. The process is inherently non-demonstrative, as Sperber and Wilson say:

> even under the best of circumstances, ... communication may fail. The addressee can neither decode nor deduce the communicator's communicative intention. The best he can do is construct an assumption on the basis of the evidence provided by the communicator's ostensive behaviour. For such an assumption, there may be confirmation but no proof. (Sperber & Wilson, 1986, p. 65).

Grice must have been well aware of the kind of difficulty raised by Harman for his picture of reasoning, particularly given that his work on meaning and communication is founded on inference to the best explanation[44]. Part of his preferred solution may have been to explore the possibility of non-demon-

[43] Like Smith, Harman also thinks that this picture of reasoning does not adapt well to practical reasoning.

[44] This kind of explanation is abductive rather than inductive: the explanations are not lawlike generalisations formed by reflecting on several instances, but propositions about the particular utterance.

strative inference rules that meet his criterion of preserving some kind of value linked to rational acceptability. In his work on reasoning he mentions non-demonstrative rules more than once (Grice, 2001, pp. 5, 6, 10, 22, 46), without going into great detail about what the rules of non-demonstrative inference might be. Such rules would differ from demonstrative ones in that an acceptable transition from true premises might produce a false conclusion. Thus reasoning could "go wrong ... through the perverseness of the world in refusing to conform to the conclusion of an impeccable non-demonstrative inference." (Grice, 2001, p. 6)[45]

I think that a promising approach, and one that can be pursued regardless of whether there turn out to be any rules of non-demonstrative inference, is to look at how deductive inference may be involved in non-demonstrative reasoning, particularly abductive reasoning. It is true, as Harman says, that in reasoning that seeks the best explanation for an observation, the conclusion of the reasoning process, the explanation, does not stand in relation to the observation or fact explained as the conclusion of a logical argument does to its premises. This shows that deductive inference cannot be all there is to non-demonstrative reasoning, but it does not establish that deductive inference plays no role in non-demonstrative reasoning. As Sperber and Wilson say, "By its very definition a non-demonstrative inference cannot *consist* in a deduction" but that leaves open the possibility that a non-demonstrative inference can contain a deduction "as one of its sub-parts" (Sperber & Wilson, 1986, p. 69). Indeed deductive reasoning ability may be central to abductive reasoning since in abductive reasoning, the explanation found, taken together with background knowledge, should logically support the observation it is supposed to explain.

Instead of looking for value-preserving non-demonstrative rules that generate an explanation from the observation that it is meant to explain, one can develop a picture of non-demonstrative reasoning according to which non-demonstrative reasoning is divided into hypothesis formation and hypothesis testing or confirmation. Deductive inference might play a role in hypothesis

[45] This is because non-demonstrative inference rules do not guarantee that value is preserved: "inference rules ... pick out transitions of acceptance in which transmission of satisfactoriness (including where appropriate truth) is guaranteed or (*in non-deductive cases) to be expected*." (Grice, 2001, p. 22, my italics).

formation or hypothesis checking or both. Sperber and Wilson outline a theory of spontaneous non-demonstrative inference of this type (1986, pp. 69–70, 108–117).

Theories in which non-demonstrative reasoning is divided into two parts are relatives in psychology of Popper's hypothetico-deductive theory of scientific discovery (Popper, 1959). However, in important respects the psychology of non-demonstrative reasoning differs from Popper's theory. Popper, whose interest was the *logic* of scientific discovery, can say of hypothesis formation that it is a non-logical, psychological process and largely leave it at that, concentrating on the logic of hypothesis testing. For a cognitive scientist, the processes involved in hypothesis formation are to be explained, just as much as the processes involved in testing hypotheses. Moreover, in areas of cognition that are typically fast and automatic, not much testing may occur, so the burden of explanation is shifted towards the account of hypothesis formation.

On either account, hypothesis formation is guesswork, partly a matter of intuition and inspiration, but for Sperber and Wilson it is "suitably constrained guesswork" (Sperber & Wilson, 1986, p. 69), and part of what constrains it is the use of deductive inference rules: "Deductive rules, we will argue, play a crucial role in non-demonstrative inference... Hypothesis formation involves the use of deductive rules, but is not totally governed by them." (Sperber & Wilson, 1986, p. 69) The idea is that new information, a set of propositions **P**; interacts with information from the context, including background knowledge and assumptions, a set of propositions **C**; to generate a contextual implication *Q*. **P** and **C** taken together logically imply *Q* since *Q* is generated from **P** and **C** by standard deductive inference rules. *Q* does not follow from either of **P** and **C** individually. Thus *Q* is not demonstratively inferable from the new information, **P**, but it is arrived at because this new information is processed according to deductive rules (in the context of background information). This process can be seen as non-demonstrative inference from **P** to *Q*, and as the hypothesis-formation stage of non-demonstrative reasoning.

This is the pith of Sperber and Wilson's account of non-demonstrative inference. There are further important details I have not explained here, some of which I go into below in the section on mental logic, and some of which I return to in chapter 4, in discussion of utterance interpretation.

A key difference with a hypothetico-deductive theory of science is that Popper's theory concerns non-demonstrative reasoning (primarily) in the context of scientific hypotheses. In the present chapter I am commenting on reasoning in general, but my interest in reasoning in this thesis is focussed on its involvement in utterance interpretation. The ways that hypotheses are generated may vary from domain to domain and it may be much harder to come up with hypotheses in some domains – scientific investigation of nature is the obvious example – than in others, such as utterance interpretation. In some domains, and here scientific theorising is a paradigm case, conscious use of rules of thumb for discovery – e.g. "try to imagine the simplest possible system with the properties that are of interest" – may commonly be part of the process (heuristics in one sense of the word: see chapter 3), but no guarantee of success. In other, limited, domains, it may be that true hypotheses are more likely to come to mind than false ones, or that, as Sperber and Wilson write, "of the assumptions that come most spontaneously to a human mind, those that are true are more likely to be [or seem] relevant than those that are false". (Sperber & Wilson, 1986, p. 117) This would mean that if the cognitive system judges a conclusion to be relevant then that is an indication, *post hoc*, that the assumption or assumptions which led to it are likely to be true.[46]

In domains in which humans are disposed to have good hunches, hypothesizing is less likely to feel effortful and laborious. Utterance interpretation plainly falls into this category. In typical cases the hearer is not aware of any of the working out that underlies his interpretation of the utterance. Still there must be mental activity involved: just because something is below the waterline does not mean it is not there. I think that a significant part of the explanation lies in what Herbert Simon calls *recognition*. Through considerable experience one's cognition comes to be set up so that when one encounters a new situation which is similar to previous ones, in conversation, as in chess (one of Simon's examples), relevant facts are automatically brought to mind on that

[46] In the same way, and for the same reasons, that corroboration of a scientific hypothesis lends support to the assumptions that it rests on.

basis[47]. Part of this picture is that it is more likely, when there is a good fit between cognition and the domain, that relevant assumptions come more readily to mind than irrelevant ones. Another part of this picture may be heuristics that jump from observations to hypothesised conclusions, or from situations to judgments. Such heuristics blur the line in interesting ways between hypothesis formation and non-demonstrative inference rules.[48]

I have sketched out a way, derived from Sperber and Wilson, that non-demonstrative inference can be split into hypothesis formation and hypothesis testing, and that hypothesis formation can involve rules of deductive inference. Sperber and Wilson see hypothesis confirmation, conversely, as a non-logical process: a process not involving rules of deductive inference. They say that confirmation of hypotheses is "a by-product of the way assumptions are processed, deductively or otherwise." (Sperber & Wilson, 1986, p. 69) As already explained, the idea is that an assumption gains support from *post hoc* strengthenings. If an assumption turns out to be fruitful then it is strengthened, because the chances are that an arbitrary assumption would not have led to interesting or useful results.

I think that whether one sees the construction of the chain of deductive inferences as part of the hypothesis-formation stage or as an aspect of hypothesis testing may depend on one's point of view as a theoretician: that is, whether it is the assumption that is regarded as the hypothesis, or the assumption together with contextual conclusions. In cases where what is of primary interest is the assumption, rather than the contextual conclusion, one might argue that hypothesis formation is limited to the non-logical process of constructing or retrieving an assumption, and that the deductive processing that follows is part of hypothesis testing. In either case, I have outlined in this section a way that non-demonstrative inference can involve value-preserving transitions.

[47] Note that while Simon calls this phenomenon *recognition*, there is no requirement that a mental representation be formed of the fact that this situation is similar to one previously encountered.

[48] This is something that Deirdre Wilson pointed out to me in discussion of these topics (p.c.).

2.2.3 A RIVAL VIEW OF RATIONALITY

There are influential opponents of views of rationality as the ability to make value-preserving transitions, among them Gilbert Harman and Richard Foley. We have already seen two of Harman's objections: 1) that reasoning is often not from logical premises to conclusions, an objection I have tried to deal with in the previous section; and 2) that it is wrong to suppose that reasoning is simply a matter of applying laws of entailment. The second objection amounts to the claim that definitions of reasoning like Grice's confuse rules of reasoning, which are psychological, with laws of derivation, which are logical. Harman writes:

> Logical principles are not directly rules of belief revision. They are not particularly about belief at all. For example, modus ponens does not say that, if one believes *p* and also believes *if p then q*, one may also believe *q*. Nor are there any principles of belief revision that directly correspond to logical principles like modus ponens. (Harman, 1984, p. 107)

I agree with Harman on the first point (disagreeing, therefore, with Dummett, 1973 and Hacking, 1979) that syntactic laws of logic are conceptually distinct from psychological rules of reasoning, but not on the second, that there are no reasoning rules corresponding to logical rules, as I discuss in section 2.3.

As well as the conceptual point, there are other reasons to doubt that rules of reasoning are in some way isomorphic with logical laws. One consideration is that there are plenty of rules of reasoning that do not resemble laws of natural deduction. For example, there may be rules of reasoning which should be applied when two or more beliefs are inconsistent. One possibility is: *abandon the belief which is less (or least) certain, then check to make sure that the remaining beliefs are consistent.* There are other possibilities, but the details do not matter in this connection. The point is that none of these rules is parallel to a law of natural deduction, unlike a rule for reasoning such as *if you believe something of the form* if p then q, *and you believe* p, *then you should conclude* q.

Harman suggests another reason why rules of reasoning cannot simply be read off from laws of logic: that there would be an explosion of inferences in any reasoner who tried to work through all logical entailments of all of his beliefs. As he says, even if one can validly deduce a conclusion from beliefs held and that conclusion is not in conflict with other beliefs:

> there may simply be no point to adding it to one's beliefs. The mind is finite. One does not want to clutter it with trivialities. It would be irrational to fill one's memory with as many as possible of the logical consequences of one's beliefs. That would be a terrible waste of time, leaving no room for other things. (Harman, 1984, p. 108)

I think that this objection becomes serious and interesting in the context of a theory of cognition that is realist about mental representations and in which rules are applied automatically if their input conditions are met. If there is no clear account of what it is for the mind to have a belief or form a new one, then it is not clear how costly it is to do so, or indeed that it is costly at all in the theory[49]. And if the rules only amount to advice about good reasoning, then the problem does not arise because they need not be followed mechanically[50]. I examine a version of this argument in the section below on mental logic, where the criteria for this to be a serious problem are met.

Provisionally setting aside, then, the question of whether there is a rule of reasoning corresponding to each syntactic law in logic, I agree that reasoning is not simply a matter of following truth-preserving steps. This point goes beyond what has been said (in section 2.2.2) about the division of non-demonstrative reasoning into hypothesis formation and hypothesis testing and the role of deductive rules in either phase. There is more to reasoning than either deductive inference, or non-demonstrative inference construed as postulation and confirmation of assumptions. Reasoning involves the abilities to detect

[49] Harman makes a distinction between explicit and implicit beliefs. It is explicit beliefs that one should not multiply needlessly, on the assumption, "that there is a limit to what one can believe explicitly" (ibid). With a realist view of mental representation and a view of reasoning as computation over these representations, it is natural to say that while there are limits on how many beliefs can be stored, there are much more strict limits on the time and effort that is available for processing.

[50] If the rules are supposed to have a strong normative force, i.e. one should not fail to follow them or one may correctly be judged irrational, then the problem would be that this is an unrealistically strong, unbounded set of normative requirements. That would not show that there is anything wrong with the rules construed as descriptive of some of the capabilities of the reasoning system.

and resolve contradictions or weaknesses, the ability to see things from a new perspective, some ability to make guesses or to hypothesise, and perhaps other abilities too. What is more, episodes of reasoning involve not only the exercise of these abilities, but much else besides. However, I think that one can concede all of this without agreeing with Dancy's contention that "Grice is wrong to link rationality so directly to inferential competence" (Dancy, 2003, p. 277)[51].

It is intuitively plausible that a great deal of what we do when we reason is not a matter of making value-preserving steps. The theoretical points made by Harman and others, as well as consideration of real examples of reasoning, suggest that episodes of reasoning have what we can call extra-logical features. Grice gives the example of someone who has agreed to give a series of lectures and is asked for the titles of the individual lectures before he has begun to think about the series. He may do a number of things which do not fit the simple version of Grice's model: think of cancelling the lectures, remember similar previous occasions, panic, decide to try to write at least the first four lectures before the course starts and so on (Grice, 2001, p. 18).

Such considerations motivate a broader view of rationality, competing with Grice's conception, as Morton explains:

> Readers of Harman or Foley will be very sceptical that in believing p one acquires a commitment to believe consequences of p, even conditional on holding on to p. Sometimes one should and sometimes one should not, depending on many factors. To be rational and intelligible is to try to revise one's beliefs in the right ways, to be sure, but these right ways are subtle and extremely hard to describe. (Morton, 2006, p. 779)[52]

Similarly Harman characterises reasoning as a kind of 'change in view' which aims at coherence and simplicity (Harman, 1999, ch.s 1 & 3). It is easy to agree

[51] Dancy's remark is made more in the context of practical reasoning than theoretical reasoning. I do not take a position on whether practical reasoning ability is, at base, the ability to make transitions that preserve practical value.

[52] A similarly broad characterisation of reasoning is given by Stenning and Monaghan but in terms that are realist about mental representation:

> Reasoning happens when we have representations of information about some situation, and we transform those representations in ways that lead them to rerepresent information about the same situation. (2004, p. 132)

with such claims, while noting that they are compatible with a Gricean/traditional picture since they are much broader or less specific, and that a Gricean picture would be preferable if found to be tenable precisely since it is more specific. Of course, views of this type are only compatible with a Gricean picture of reasoning if 1) they are not meant in the strong sense that reasoning is not mainly, or at all, a matter of constructing or rehearsing arguments, and if 2) it is possible to relax Grice's picture somewhat, so that the ability to make value-preserving transitions, while at the heart of reasoning ability, does not exhaust it. Grice saw the need for just such a broadening of his picture of rationality.

2.2.4 FLAT AND VARIABLE RATIONALITY

Grice's suggested solution is to distinguish between a basic notion of flat rationality – "the capacity to apply inferential rules" (Grice, 2001, p. 27) – which any agent who can reason to a certain minimal standard[53] would possess by definition, and a variable notion according to which agents who reason better are more rational than others. The concept of variable rationality would then be derivable from the concept of flat rationality together with the fact that flat rationality is used to solve problems. (The goal-directed nature of reasoning – which has already been mentioned – is examined further in the next section). Grice also considers, but decides against, the possibility that the concept of flat rationality is derived as a limiting case of variable rationality. I explore the distinction between flat and variable rationality, then show how it can be put to use in dealing with objections like Harman's.

As noted in chapter 1, Grice's distinction somewhat resembles the competence/performance distinction in modern linguistics, although Grice calls flat rationality a 'capacity'[54] and variable rationality 'a competence'. It also resembles a distinction that Christopher Cherniak (1986) makes between minimal descriptive and minimal normative rationality. Minimal descriptive rationality is the threshold level of reasoning ability that must be possessed by

[53] Or 'Rational Being' in Grice's terminology.

[54] On p. 27 (see the quotation in the previous paragraph). On p. 30 Grice talks of an "unfailing competence with respect to certain rudimentary inferential moves." See the discussion of this point in this section, below.

any rational agent, by definition (i.e. *qua* rational agent). Minimal normative rationality sets a level that rational beings should aim for. The similarity is that if someone falls short of Cherniak's minimal normative rationality, or possesses little of Grice's variable rationality, perhaps overlooking a relevant and straightforward inference because of confusion, then that person is not as rational as he should be. People who behave this way, and behaviour of this kind, are often called irrational. On the other hand, an agent who fails to make any simple inferences even in favourable circumstances is apparently not capable of rational thought. In Cherniak's terms a being not capable of any inferences falls short of minimal descriptive rationality; in Grice's it fails to exhibit flat rationality and would not be a rational being. A being[55] of this sort is more aptly called a-rational or non-rational than irrational.

Grice calls variable rationality an 'excellence' as well as a competence. Variable rationality is something that it is good to possess more of, if you are rational at all. In the same way it is good for any minimally rational agent to approach normative rationality. According to Grice, a person possessed of a high variable rationality quotient would have strengths in areas not strictly necessary to reasoning, but helpful for good reasoning. Grice mentions several such properties, including clear-headedness, a sense of relevance, flexibility, inventiveness, thoroughness and 'nose' (intuitiveness) (Grice, 2001, p. 31, including footnote 3). An interesting comparison can be made with desiderata for good reasoning that are sometimes given in the context of teaching people to reason better. A typical example is the list of "Abilities, qualities and propensities that good reasoners are likely to possess" provided by Nickerson (2004). This list includes several kinds of knowledge and motivational factors, including domain-specific knowledge, self knowledge, a strong desire to hold true beliefs, and curiosity/inquisitiveness (Nickerson, 2004, p. 415), all of which, to my mind, are still further from the core of reasoning ability than the properties Grice lists. I think that some aspects of variable rationality, particularly those on Grice's list which concern intuition and instinct, may be best explained in terms of the ways that a cognitive system is well tuned to the domain (or domains) in which it operates. As mentioned above, intuition in complex tasks may be partly a matter of having enough experience in the rel-

[55] Cherniak's view seems to be that such a being is neither an agent nor rational.

evant domain so that suitable knowledge is quickly brought to bear. The other side of the same coin is that good, quick performance depends on ignoring the vast majority of potentially relevant information. I comment further on this in the section on heuristics in chapter 3. Other aspects, such as the degree and extent of someone's motivation, seem to me likely to lie beyond what it is possible to investigate scientifically at present.

We now have a refinement of Grice's original suppositions: rationality is split into a core capacity for value-preserving inference plus more peripheral attributes relating to intuition, and perhaps also still more peripheral features to do with motivation or knowledge of certain domains. Does this view allow good responses to Harman's criticisms? The criticisms are rooted, I think, in an observation and a conceptual point. The observation, which is also Grice's starting point for complicating his picture with a distinction between flat and variable rationality, is that episodes of reasoning involve much more than following inference rules. The conceptual point is that logical laws and psychological rules are quite different types of thing. I think that the conceptual objection can be granted but set aside. No doubt it is true that there is a difference between logical laws, with a status similar to laws of mathematics, and rules of the working of the mind. However, that does not exclude the possibility that some psychological rules are isomorphic with rules of logical inference in the way proposed by the traditional view of reasoning. It is an empirical matter to find out which, if any, psychological rules for reasoning are isomorphic with laws of deduction and to investigate which other psychological rules or procedures there are to augment them or to shortcut them: for example, rules for resolving contradictions, for building mental models and perhaps for shifting focus.

Turning to the observation that episodes of reasoning can involve much more than value-preserving transitions, the separation of flat from variable rationality has the advantage of explaining how this can be, while reasoning ability is essentially the ability to effect such transitions. We can see that a reasoner with excellent variable rationality in addition to the basic flat capacity might have made less of a meal of the lecture crisis, remaining focussed on the task of deciding what to do, in the light of the situation and his aims. Another reasoner might be intimidated by the situation but still quick to intuit ('nose out') a solution. The important point, in the context of the criticisms made by Harman, Dancy and others, is that there is no difficulty in seeing that a basic capacity for value-preserving inference is compatible with the possession or non-possession of intuitions, motivation to reason and abilities such as those for resolution of contradictions and recovery from contradiction.

Returning to the concept of 'flat rationality', endorsing the view that flat rationality comes before variable rationality (in order of derivation), has the consequence – congenial to a cognitive scientist, although Grice saw it as embarrassing – of committing oneself to the view that "there is a specifiable minimal competence held by all RBs [rational beings]... [this view] seems to involve attributing to all rational creatures, as the core of their rationality, an unfailing competence with respect to certain rudimentary inferential moves." (Grice, 2001, p. 30)[56]

There is a kind of ambiguity in this statement (hinging on the scope taken by 'certain'). Grice might mean that any rational being must have a competence in some inferential moves or other, perhaps different ones for different rational beings, with no common core of inferential moves that any being must possess in order to count as a rational being. Perhaps it is more plausible that what was meant is that there are some particular inferential moves, the possession of which is a necessary condition of rationality. (The use of 'certain' rather than 'some' suggests this was Grice's intended meaning, I think, as does the analogy he makes with chess-playing ability[57].)

Grice does not discuss this aspect of the core of rational competence further, but there is discussion of just this point in Cherniak (1986, chapter 2) (without reference to Grice). Cherniak rejects the view that there are particular inferential moves which all rational beings, as rational beings, must be competent in, but endorses the weaker thesis that a rational being must be capable of making some inferences. An agent with an "inverted feasibility or-

[56] Either way, it might be that what Grice found embarrassing was that the competence would need (or so he thought) to be unfailing (Deirdre Wilson, p.c.). I do not think that the priority of flat rationality over variable rationality would entail infallible *performance* with any inference rules.

[57] Someone who does not know how each type of piece moves in chess does not know the game, and cannot be called a chess player, even if, improbably, he were to have some grasp of higher level principles such as tactics in the endgame.

dering" of inferences – inverted relative to human abilities, that is – might find it easier to infer $\forall xFx \rightarrow \forall xGx$ from $\exists x\forall y(Fx \rightarrow Gy)$ than to perform *modus ponens* (Cherniak, 1986, p. 34). Is an agent of this type possible? Cherniak argues that such an agent could exist. A being that has no memory for theorems that it has derived would have to make all inferences by reference to a static body of knowledge, a deductive system of axioms and rules. Any inference corresponding to one of the rules or axioms would be easy for this agent, regardless whether the rule looks complex to us, such as the inference rule $(\forall xFx \rightarrow \forall xGx) \vdash (\exists x\forall y(Fx \rightarrow Gy))$ or is an apparently simple one like $P \rightarrow Q, P \vdash Q$.

I think that, accepting a Gricean picture of rationality, it is hard to avoid the conclusion Cherniak reaches, that a being is rational as long as it has a core competence with some truth-preserving transitions, and there could be rational beings which would surprise us by failing to make inferences that we find obvious, such as *modus ponens* or and-elimination. Finding that a being does not have some particular inference rule in its repertoire does not, I think, justify the conclusion that the being in question is non-rational. This may be just as well, since the evidence is that *modus tollens* is not psychologically basic for human beings (see §§2.3 & 3.2). Related evidence supports the conclusion that human beings possess a small set of deductive inference rules, although it is a matter of debate which ones. In my view it would be surprising if the deductive inference rules available for spontaneous inference differed greatly from person to person, so I think that it is true that there are particular inferential moves that all humans are competent in, barring pathology, even if some other – imaginary – rational beings could be rational in virtue of competence in a different set of basic inference rules.

2.2.5 THE HARD WAY AND THE QUICK WAY

In the introduction to this chapter I said that one aim was to show the compatibility of a Gricean view of rationality with the apparent psychological reality that reasoning makes use of shortcuts and heuristics: that, in other words, reasoning is not fully explicit and the transitions made may be unsound. Here I explain what I mean by 'heuristic' and 'shortcut' and set out the conflict with the traditional view of reasoning. Full exploration of the evidence for heuristics and shortcuts in reasoning and of the varieties of heuristics and shortcuts in use is left for the next chapter. In this section I use examples that Grice provides to illustrate the point that reasoning is not fully explicit.

It appears hard to reconcile the Gricean view with research that shows that a good deal of cognition involves heuristics. Heuristics are 'rules of thumb': rules that work well enough most of the time in their intended domain but do not invariably produce correct output from correct input. Some heuristics are like inference rules in that they perform transitions from input information to conclusions[58], but unlike them in that they do not guarantee that the transitions are value-preserving. An inference performed by such a heuristic is, by definition, unsound, in contrast to the inferences performed by according to the rules we have been considering, which are sound, also by definition. As an example consider the rule of thumb that birds can generally fly, which could be used to make an unsound but often correct inference from *x is a bird* to *x flies.*

Some other shortcuts are value-preserving, but share with heuristics the property of inexplicitness. Both skip over steps that a fully explicit derivation would include[59]. It is intuitively plausible that much of reasoning is performed by heuristic or otherwise not fully explicit shortcuts, as some examples given by Grice suggest. Evidence from psychology of reasoning also provides strong support for this view.

Thus there appears to be a clash with the traditional or Gricean picture set out so far. This is a different issue from the observation that a good deal of reasoning involves much more than chains of deductive inference, discussed in the previous section. The problem here is rather that if reasoning often or typically involves shortcuts, then it often or typically does not involve steps that parallel logical inferences, and this is apparently in direct contradiction to the model of reasoning that I am advocating.

[58] Some heuristics work at a different level, regulating which procedures are followed, rather than performing (or mandating the performance of) particular transitions.

[59] Roberts (2004) suggests that the psychologically more interesting category is shortcuts rather than heuristics. The idea is that what matters is for psychology is the distinction between fully explicit reasoning and reasoning in which some steps are omitted. This distinction is important, but for my purposes at least, the distinction between sound and unsound rules is important too.

Grice did not think that reasoning in general was always explicit. He gives the example of a six page proof or sketch of a proof by Georg Kriesel (whom he gave the pseudonym *Botvinnik*) which was later expanded to a more complete proof of eighty-four pages. The long proof contains many steps that are left implicit or glossed over in the original. Consideration of this and simpler examples makes it clear that we skip steps as far as our conscious train of thought is concerned. It is not just that Kriesel did not write out all of the steps of the expanded proof: it is highly unlikely that he was aware of all of them even as he worked the proof out.

We might try to extend Grice's model to incomplete reasoning by saying that examples such as Kriesel's proof are reasoning (or good reasoning) because one could complete them by supplying extra premises to make a deductively valid argument. This also works for simpler cases: if I reason from *Jack is an Englishman* to *Jack is brave*, we can make the reasoning complete by supplying as a missing premise, *All Englishmen are brave.* (Grice, 2001, pp. 8–10).

An argument with a missing premise is traditionally called 'enthymematic'. Most arguments presented in speech are enthymematic, presumably because it would usually be pointless to outline premises that the hearer can infer for himself[60], and, given that fact, counterproductive to do so, because it suggests to the hearer that there was a reason for spelling out the premise explicitly, and this is liable to send him off on the wrong track in interpreting the inference. (On arguments incomplete in this way,.

The proposal is that an argument which is enthymematic or otherwise incomplete is informally valid if and only if there is a complete argument which is valid and is identical to the incomplete argument except for the addition of propositions. The extra propositions supplied may be premises, intermediate stages in the argument, or even conclusions.

There is, however, a problem with this way of rescuing the traditional picture. Almost any sequence of propositions can be seen as a valid argument if one is allowed to supply additional premises without constraint. (On this

[60] Davidson (1963) makes a similar point about the reasons we give for actions : e.g. *I pressed the switch because I wanted to turn the light on.* There is usually no need to add: *and I believed that pressing the switch would turn the light on.*

point, see Mill, 1856, p. 527.) So this criterion would make it impossible to distinguish between good reasoning and bad reasoning or non-reasoning. Grice gives the example of Shropshire, a budding philosopher who claimed that the fact the chickens run around after decapitation proves the immortality of the soul. It is not clear that one would want Shropshire's performance to count as reasoning, but one can expand his two-proposition sequence into a sequence that is canonically valid, as Grice demonstrates (2001, p. 11). (See Appendix I for Grice's expansion of Shropshire's argument.)

One way of making the desired distinction would be to suppose that in reasoning that appears to be incomplete, although the individual steps are not spoken, written or *consciously* entertained, they are part of the mental process by which the conclusion is reached. Then the distinction between reasoning or non-reasoning would be that in reasoning all of the steps are present, either consciously or subliminally, whereas in non-reasoning, there is no complete chain of steps in the mind between the premises and the conclusions. In good reasoning the steps would be value-preserving. Bad reasoning might involve steps that are presented as, or thought to be, value-preserving but are not. However, I do not think that this amendation is promising because it is implausible that all steps in reasoning are explicitly made, whether consciously or otherwise. Before explaining this criticism, I consider another criticism that I do not find convincing.

I take it that the following complaint about inferentialism, made by Audi, is aimed at theories of reasoning of the traditional kind:

> One error [that philosophers make in accounting for rationality] is inferentialism: the tendency to posit far more inferences than we usually make or – unless inference is reduced to a mere brain process as opposed to a mental operation – even *can* make in the rational conduct of our lives. (Audi, 2001, p. viii)

There is an implication that the upper limit on the amount of inference that we can perform is known. It may be that Audi is relying here on introspective evidence: we do not, even when we introspect carefully, find that we are aware of performing all the inferential steps that would be part of completely worked out versions of complex arguments. This is not a decisive objection, however, since, as suggested above, these steps might be going on without ('beneath') our awareness. It is a commonplace in psychology and cognitive science that introspection is not always reliable, and another commonplace in these subjects and in linguistics that we are not aware of all the processes that go on in our minds, that some processing is, in fact, completely inaccessible to conscious introspection. An objection based on the fact that we do not, when we introspect, find that all the steps posited by inferentialism are available, would need to be coupled with an argument that such steps, if they happen at all, must (unlike linguistic or visual processing) be introspectable. Recanati has offered an argument of this kind as part of his reason for distinguishing between non-inferential primary pragmatic processes and genuinely inferential secondary processes. I consider his views in chapter 4.

I do not want to advocate this kind of inferentialism, however, for a different reason (which may have been Audi's reason too). I think that considerations of time and processing effort make it clear that the mind must use shortcuts during inferential processing rather than spelling out each value-preserving step of each inferential chain. Since I do not think, however, that fully explicit inferential steps are necessarily conscious, I think that the two issues are orthogonal. That is, I think that there are two separate questions: (1) whether inferences are performed by fully explicit truth-preserving rules or by shortcuts; and (2) whether the mental steps involved are conscious or not. I think, in fact, that all four logical possibilities are instantiated: (1) conscious, fully explicit reasoning; (2) conscious use of shortcuts in reasoning; (3) reasoning which involves mental representation of each step of a logical deduction but in which some steps are not conscious; and (4) reasoning which is inexplicit in that it skips steps and is also not consciously available. I give examples of each type in chapter 3, where I discuss heuristics at greater length.

If at least some reasoning involves shortcuts which are not filled in explicitly 'behind the scenes' in the mind, then the proposed refinement to the Gricean picture will fail to distinguish in a principled way between good reasoning and bad or non-reasoning. Another way of making the distinction is needed. Grice provides two suggestions, both involving the intention with which reasoning is performed. One proposal is that incomplete or informal reasoning counts as reasoning because it is intended to be value-preserving:

> we could say ... that x reasons (informally) from A to B just in case that x thinks that A and *intends* that, in thinking B, he should be thinking something which would be the conclusion of a formally valid argument, the premises of which are a supplementation of A (Grice, 2001, p. 16, his emphasis).

The second proposal brings us back to shortcuts and heuristics:

> "We have... a 'hard way' of making inferential moves; [a] laborious, step-by-step procedure [which] consumes time and energy... .A substitute for the hard way, the quick way, ... made possible by habituation and intention, is [also] available to us, and the capacity for it (which is sometimes called intelligence and is known to be variable in degree) is a desirable quality". (Grice, 2001, p. 17)

The important point here is that the 'quick way' of making inferential moves counts as reasoning in Grice's model. Grice is quite clear about this[61]. Reasoning ability is centrally the ability to perform valid transitions between thoughts, but the transitions need not all be explicitly spelled out in any given episode of reasoning. The idea is that if a certain transition (or kind of transition) is made repeatedly, then a shortcut may be found. In future reasoning the shortcut is used with the intention that it leads where fully explicit steps would have led: to a valid conclusion.

Grice's two proposals have in common this appeal to intentions. This appeal is also partly motivated by the fact that there are canonically valid strings of inferences that are not good examples of reasoning. What such examples share is that the inferences in these strings do not seem to be directed, as the following example illustrates:

> Suppose ... that I were to break off the chapter at this point, and switch suddenly to this argument: "I have two hands (here is one hand and here is another). If I had three more hands, I would have five. If I were to have double that number I would have ten, and if four of them were removed

[61] In his introduction to (Grice, 2001), Warner concurs: "Kriesel's quick way leaps over the vast majority of [the] steps [in the complete proof], but it is still reasoning, still an exercise of the ability to make reason-preserving transitions." (Warner, 2001, p. xxxiii)

> six would remain. So I would have four more hands than I have now." Is one happy to describe this performance as reasoning? There is, however, little doubt that I have produced a canonically acceptable chain of statements. (Grice, 2001, p. 16)

Examples like this suggest that a string of inferences, even one that is canonically valid and complete, is not reasoning (or only barely so) unless it is somehow directed. Aimless inferring will not do. Conversely, the production of a sequence of propositions that is going somewhere and which is related in the right way to a canonically valid sequence is reasoning:

> A mere flow of ideas minimally qualifies as reasoning, even if it happens to be logically respectable. But if it is directed, or even monitored (with intervention should it go astray, not only into fallacy or mistake, but also into such things as irrelevance), that is another matter. (Grice, 2001, p. 16)

It is an extension of this point to suggest that the production of incomplete sequences of thoughts is still reasoning if it is accompanied by the right intention.

In order to make the pervasive use of shortcuts (some of which are merely heuristic) compatible with the traditional picture of reasoning, it seems we must adopt something like Grice's modification of the picture, according to which a flow of ideas completely isomorphic with a logical derivation is neither sufficient nor necessary for reasoning, since it is necessary to have the right intention, and some incomplete sequences of thoughts are also reasoning, if accompanied by suitable intentions.

Bringing in intentions might seem to be a dangerous manoeuvre, making our picture of reasoning dependent on the resolution of difficult issues in the philosophy of action, where a great deal of attention has been paid to the connections among beliefs, desires, intentions and actions (e.g. Davidson, 1963; Bratman, 1987; Mele, 1997b). Here I look at one problem that arises and suggest that a better solution might be to say, more neutrally, that inference must be goal-directed to count as reasoning.

One apparent problem with the view that shortcuts count as reasoning because they are intended to preserve the truth of the premises, is that when reasoning is quick and inexplicit it is unlikely to be accompanied by any conscious, explicit intention that the output is correctly related to the input. We would need to say, then, that the intention can be implicit. So this claim about reasoning would depend on the view that one does not need to be saying to oneself, 'My intention in doing *x* is to accomplish *y*', or to be conscious that one has intention *x*, or even that one is attempting *y*, to be truly said to have intention *x*. This might be acceptable – certainly one can do something intentionally without any conscious intention to do whatever it is. However it is not clear that doing something intentionally necessarily involves intending to do it. (This is what Michael Bratman (1984) calls the 'Simple View'. It has been much debated. Nadelhoffer, 2006, is a useful recent summary.) These are difficult and controversial issues[62]. Fortunately, I think that there is no need to resolve them here. There is a further consideration against bringing in intentions.

For some types of inference that I would like to regard as reasoning, it seems that there need not be any intention to reason correctly or to achieve a certain goal. In the case of hearing an utterance it is, at best, odd to say that one intends to understand or intends to try to understand it. It is very odd indeed to say that the inferences involved in working out what a speaker meant were accompanied by an intention that they be valid inferences, since a hearer is not typically aware of making any inferences at all in understanding an utterance.

A reason for the oddity might be that intentions, or talk about intentions, is at what is sometimes called the person level. When we talk about intentions we see them as properties of a person. The reasoning that I am concerned with, however, particularly when it is quick and subliminal, seems to be conducted by mental subroutines: dedicated modules or processes. Can an intention also be effectively a property of a module, strategy or process? I would rather not say that. Can we say, instead, that a module has a goal or a function? Stanovich and West (2004) say that it cannot:

[62] On the 'Simple View', much of the debate has been about folk psychology or about the meanings of the words involved, not always clearly distinguished, rather than about the role of mental states in behaviour. See Nuti, 2003 for distinctions between the study of psychology, of folk psychology and of semantics in discussions of *belief*, *intention* etc.

> we do not think the question of whether a certain (internal) strategy is rational or irrational is well formed. We do not believe the term rationality applies to subpersonal entities. ... One could ... talk of a submodule that chose strategies rationally or not. ... [the question would arise] what are the goals of this subpersonal entity – what are its interests that its rationality is trying to serve? This is unclear in the case of a subpersonal entity. (Stanovich & West, 2004, p. 532).

My suggestion is that it does make sense to talk about the goals or functions of at least some 'subpersonal' processes. These goals are effectively hardwired into their structure, either by evolution, or by learning. A bicycle-riding module would be an example of a learned module with a learned function; an utterance interpretation module is apparently an innate module with a function given innately, as is the language-parsing module. The bicycle-riding module consists of procedures that serve the dual purpose of getting the rider where he wants to go while keeping him on the bike. An utterance interpretation module has the purpose of constructing correct interpretations of utterances.

Thus, in ordinary language terms, in normal circumstances it is strange to say that *I* try to understand an utterance: I find myself understanding it, or not, and if not, then I might subsequently try to understand it by ruminating or seeking further information. On the other hand, my pragmatics module can be said, loosely speaking, to try to find a correct interpretation for an utterance. It has (speaking less loosely) the function of assigning interpretations to utterances: its goal on receiving input relating to an utterance is to arrive at the correct interpretation, or one that is near enough to correct for current purposes. It is plausible that the module obeys certain regulatory principles which can be seen as directing deductive steps and heuristic processes to that end. One principle that would seem to be essential, if utterance interpretation is inferential at all, is that inferential steps taken are generally value-preserving.

My proposal is that a goal-directed sequence of steps will count as reasoning, even if the steps do not entirely mirror a logical deduction, just as long as the steps are directed towards being value-preserving. There may be many central modules and procedures which meet this requirement. In chapter 4, I return to consideration of utterance interpretation along exactly these lines.

Not all modules or mental procedures with functions perform reasoning, though, since they do not all perform inference. The input to some, particularly modules for perception, is not in a suitable form for inference. Those that do perform inference, however, can be seen as performing reasoning – in the sense of inference that involves, at its core, transitions that aim at being value-preserving, in pursuit of a goal. I say more later in this chapter and in chapter 4 about the distinction between real inference, performed by central (conceptual) processes, and pseudo-inference, performed by peripheral (non-conceptual, often perceptual) processes. If this distinction, and the notion of functions of some sub-personal cognitive components, can both be sustained, as I think they can, then Grice's view of reasoning as goal-directed inference will accommodate the utterance interpretation module.

There are some caveats that need to be mentioned. One is that this way of defining reasoning allows us to include inferences accomplished largely by heuristics, but it will only work for some transitions involving heuristics. It will not work, for example, for conscious use of heuristics that are known to the user to be so inaccurate that they could not be intended to be value-preserving.[63] That is as it should be, I think. Making judgements by deliberate use of a rule of thumb that one knows full well consistently fails should not count as reasoning.

On the other hand, heuristics that are accurate within a domain will fit the revised definition. One can certainly rationally intend to use such a heuristic to reach a canonically correct (i.e. value-preserving) answer within a domain to which it is well-fitted, since within that domain, the heuristic generally is value-preserving. Similarly, such a heuristic can be said to be serving the function of a module or process within that domain.

A second caveat is that I have written as though intentional behaviour and goal-directed behaviour can play a similar role in a definition of reasoning, except that intentions are plausibly only attributable to people, whereas mental modules or processes can have functions that direct them towards goals. There is another difference, however. Some intentional behaviour is not direc-

[63] Conversely, manoeuvres of this kind may not be needed where procedures use shortcuts that are in fact value-preserving or algorithmic. I look at some examples of algorithmic shortcuts in chapter 3.

ted towards any goal beyond itself. Things that one might intend – or do intentionally – with no goal beyond doing them include whistling in the kitchen and drinking a can of paint (the former example is from Mele (2001, p. 28), the latter is Davidson's (1963)). I do not consider this question here.

2.2.6 THEORETICAL AND PRACTICAL RATIONALITY

As noted in the introduction to this chapter, it is usual in philosophy to make a distinction between theoretical and practical rationality, where "theoretical rationality is concerned with what to believe ...[whereas] practical rationality is concerned with what it is rational to do or to intend or desire to do." (Mele & Rawling, 2004a, p. 3) The former is the rationality exhibited (or not) by beliefs or by the process of arriving at a belief, or by a person insofar as his beliefs are rational. The latter is the rationality applicable to actions or the intentions to perform actions, or to the process involved in arriving at intentions, or, again, to a person whose intended actions are rational. Practical reasoning ability is the ability to respond to practical reasons, that is, reasons to do or to intend to do something. These are generally thought to be supplied by beliefs and desires or plans. If I have a fixed plan to improve the appearance of my neighbourhood and I believe that mowing the lawn will help bring that about, then I have a reason to mow the lawn. Theoretical reasoning ability, on the other hand, is the ability to deal with theoretical reasons, that is, reasons to believe something, and these reasons are to do with the support for a proposition – whether it is true or evidenced – as discussed above. The distinction is generally clear, although perhaps not in some special cases which I do not discuss here.[64]

Although it is traditional to divide rationality into practical and theoretical, finer subdivisions are certainly possible. Cohen, for example, identifies nine types of rationality, including deductive reasoning, mathematical reasoning, semantic reasoning, inductive reasoning, probabilistic reasoning, reasoning

[64] One that Harman raises is that one can have practical reasons for beliefs; so practical reasoning could lead to the possession of certain beliefs (Harman, 2004). If God punishes non-believers then one has a practical reason to believe in God, for example. This kind of complication might be dealt with by noting that those reasons are not directly reasons for beliefs, but reasons to intend to do something: to form a certain belief.

about the means required to bring something about, reasoning about the ends that action should serve, and Gricean reasoning about utterances (Cohen, 1992). Deductive and non-demonstrative theoretical reasoning have been dealt with in previous sections. Mathematical reasoning can be brought within the traditional picture as involving value-preserving transitions (Grice's 'hands' example above is of mathematical inference that is value-preserving in this way) but it is outside the scope of this thesis to expand on that claim or defend it. I comment briefly on the relation between deductive reasoning and semantic (or conceptual) reasoning in the section below on mental logic. I consider probabilistic reasoning in some detail in the next chapter.

In this section I look briefly at the prospect of bringing practical rationality under the traditional theory expounded above, then make some remarks about the reasoning involved in making (rather than interpreting) utterances.

As mentioned above, Grice proposed that practical reasoning, like theoretical reasoning, is value-preserving; here the value preserved is practical value or 'goodness' (Grice, 2001, pp. 87–88). For Grice, the common factor between the values preserved by reasoning is satisfactoriness, which really amounts to rational acceptability. Theoretical rationality preserves truth: thus if it is rationally acceptable to believe some premises then it is rationally acceptable to believe a conclusion validly derived from them. Similarly for practical reasoning, which Grice wanted to treat as preserving practical value: if there are things it is rationally acceptable to intend or to do, then practical reasoning from them should lead only to other intentions or actions that are rationally acceptable.

For Grice, this was connected with a theory of equivocality of the modal terms, words like 'ought', 'must', 'may' and 'should' as in example 6. Grice thought that the two senses of such expressions derive from a common core meaning.

(6) John should be here by now.

There is an alethic sense of the sentence, meaning something like: on the basis of the evidence available, one can infer with probability that John is here by now. This sense is analysed in (7).

(7) Acc + ⊢ + John be here by now

There is also a practical sense, meaning something like: it is rationally required – according to some rule or standard – that John is here by now. This sense is decomposed as in (8)[65]:

(8) Acc + ! + John be here by now

The assertion sign and the exclamation mark stand for 'moods' (or 'modes' – Grice altered his terminology when informed that his use of the word 'mood' clashed with the standard use in linguistics). The formula in (7) can be read as "It is rationally acceptable that it is true that John be here by now". The formula in (8) can be read as "It is rationally acceptable that let it be that John be here by now". I do not know whether the Equivocality Thesis is correct or whether an account of practical reasoning in terms of value-preserving transitions is viable.

In the remainder of this thesis I largely put aside issues concerning practical rationality. One exception is the discussion of decision theory's conception of rationality as applied to preferences (in chapter 3). I include this because decision theory is the key example of a theory where rationality is reduced to global consistency, and because standard game theory, founded in the decision-theoretic axioms, has been used to model the theoretical reasoning involved in utterance interpretation (Parikh, 1991; Parikh, 2001; Benz, Jäger, & van Rooij, 2006). (I have raised doubts about the tenability of the model in Allott, 2006.)

On the face of it, a theorist concerned with utterance interpretation needs to consider both theoretical and practical rationality. Utterance interpretation involves forming beliefs about speaker's meaning on the basis of features of the utterance that warrant such belief, and making an utterance involves having a particular kind of intention. One way of seeing this would be that inter-

[65] There has been considerable work on casting practical statements in terms of optatives, sentences of the form: *Let it be that p*, including Hare, 1952; Kenny, 1963; Goldman, 1970 as well as Grice, 2001.

preting utterances involves theoretical reasoning while making utterances requires practical reasoning.

However, the important part of the process of utterance formation for pragmatic theory to explain is not how the basic intention to convey some particular meaning is formed[66], but how the intention arises to do so by certain means: using a certain form of words, for example. The part of utterance production that is most amenable to theoretical description, I suggest, will be the part that takes for granted personal preferences and a rough characterisation of the intended meaning and explains how the speaker comes up with a particular utterance (which she thinks will convey the desired meaning). This is a rather specific kind of reasoning about means, and it comes down to reasoning about what conclusions a hearer will reach, as I explain in chapter 4.

This completes my survey of what I have called a traditional theory of rationality. In the remainder of the chapter I look at efforts to investigate rationality in cognitive science in mentally realistic terms. In the next section I look at mental logic, the theory that there are rules of reasoning parallel to the syntactic rules of logic, along with a rival, mental model theory.

2.3 MENTAL LOGIC AND MENTAL MODELS

2.3.1 INTRODUCTION

Mental logic theories propose that deductive reasoning is carried out by the operation of psychological rules that are isomorphic to laws of derivation (syntactic rules) in logic (some key works are Braine, 1978; Braine, Reiser, & Rumain, 1984; Braine & O'Brien, 1998; Rips, 1983; Rips, 1994; Rips, 1997). This programme[67] can be seen as a way of fleshing out the view that I have been presenting, that reasoning is performed by value-preserving rules. Mental logic extends this idea with a realistic view of mental representation and processing. The assumption is made that representation and processing are

[66] I assume that this question lies outside of pragmatic theory, and perhaps outside of science in general (see chapter 4).

[67] Braine and O'Brien's and Rip's theories differ but share core commitments. They can be seen as two ways to pursue the same research programme, in Lakatos' sense of the term (Lakatos, 1970).

separate aspects of the mechanics of cognition. Reasoning is then a transition or series of transitions between mental representations in working memory, where the mental representations are, like sentences in natural language or formulae in propositional or predicate calculus, sets with hierarchical structure. The transitions that are possible are those which correspond to certain syntactic rules of logic. Rips summarises thus:

> I assume that when people confront a problem that calls for deduction they attempt to solve it by generating in working memory a set of sentences linking the premises or givens of the problem to the conclusion or solution. Each link in this network embodies an inference rule ... , which the individual recognizes as intuitively sound. (Rips, 1994, p. 104)

A further assumption is made that processing is costly. An inference that involves several inferential steps will be more costly, i.e. more effortful and difficult, than one with fewer steps. If we postulate a set of basic rules, then we can make predictions about the relative difficulty of inferences. Conversely, if it is found experimentally that a certain inference is relatively difficult, then one can infer that that inference requires several steps. An inference of this form is not a basic inference accomplished in one step by using one rule from the set of mental inference rules (or 'mental logic').

This picture of reasoning should seem familiar. It is essentially the traditional picture that has been outlined and advocated above, although it does not explicitly take account of the refinements discussed above that a) reasoning may involve more than inferential transitions, and that b) reasoning may on occasions be accomplished by shortcuts that bypass inferential transitions. In my view, mental logic theory describes the core of reasoning ability. I discuss in chapter 3 the possibility that sound mental rules for deduction and heuristics can coexist.

The major rival to the mental logic programme is the theory of mental models. According to this theory, the mechanism behind reasoning is the construction of mental models of states of affairs, which yield conclusions on examination (Johnson-Laird, 1983; Johnson-Laird, 1999; Johnson-Laird, Girotto, & Legrenzi, 2003).

The input, typically a sentence or sentences, is parsed into a form suitable for the construction of mental models. Then mental models are generated by the listing of states of affairs compatible with the proposition expressed by the sentence or sentences. This stage is like the generation of truth-tables for formulae in propositional logic, but with the difference that cases that are false are not explicitly represented in the basic models.

For example, suppose that the input is the sentence, “The book is on the table or the pen is on the floor”, and that is followed by the sentence “The pen is not on the floor”. The representation for the first sentence is given in table 2:

Table 2: Mental models for inclusive disjunction (based on Johnson-Laird, 2004, p. 173)

(book on_table)	
	(pen on_floor)
(book on_table)	(pen on_floor)

Adding the information given by the second sentence rules out the models on the second and third lines of the table, leaving only the model in the first line. Thus the state of affairs described by the first model is selected as a conclusion.

Not all of the models mandated by an input need be built. Mental model theorists suggest that only one model is built in spontaneous, implicit inference, and agents do not search for alternatives unless evidence is encountered for them (Johnson-Laird, 1983, p. 127; see also Johnson-Laird, 2004, p. 188).

In some cases, reasoning may continue beyond the first conclusion reached. After a first mental model is constructed and a possible conclusion is read off the model, the conclusion can be tested. To do this, more mental models can be generated and examined. If none contradict the conclusion then it is kept; if one or more are in conflict with it, then another conclusion, compatible with all the models, is sought.

Both schools of thought claim that their theory has been well-tested and found to be supported, posing an interesting, if familiar, problem for philosophers of science. (See Evans, Newstead, & Byrne, 1993 (ch. 2) for discussion of the evidence and (ch. 3) the theoretical debate). O'Brien summarises the evidence for mental logic:

> The theory [Braine and O'Brien's version of mental logic] has predicted, successfully, which logical-reasoning judgments people make easily and which they find difficult, which inferences are made effortlessly during text comprehension, and which judgments differ from what should be expected if people were using standard logic instead of mental logic. (O'Brien, 2004, p. 205)

Equally, Johnson-Laird believes that the evidence is that the innate deductive competence of 'naive reasoners' is based on mental models (Johnson-Laird, 1999, p. 130). The claim here too is that the theory's predictions about the relative difficulty of inferences have been corroborated. A further claim is that there is evidence that people make certain unsound inferences which mental model theory predicts, but which mental logic does not. These inferences are interesting for this thesis because they show that the procedure set out in mental model theory is actually a heuristic, in the strong sense that it sometimes produces false conclusions from true premises. (I discuss this later in this section.)

The debate between the two theories is about whether the mental processing involved in deductive reasoning is similar to syntactic proofs in classical logic or to semantic proofs. These positions have in common the postulation of "universal principles for deductive competence" (Evans, Ellis, & Newstead, 1996, p. 1088) based on properties of classical logic. That is, the logical form of the input propositions and of the mental representations is what plays the crucial role in determining what conclusions are reached. Other theories of reasoning claim that the content of premises plays a role in determining what conclusions are drawn. These theories, some of which I look at in more detail in the next chapter, include the heuristics and biases programme, which postulates simple heuristics for deduction which may be sensitive to the form or content of premises and conclusions; and Oaksford and Chater's work on probabilistic optimisation, agnostic about the underlying mental representations or processes. Other possibilities which have been suggested include the theory that conclusions are selected purely on the basis of plausibility, with no regard to the support provided to them by the premises; and deduction carried out according to a mixed bag of strategies (Schaeken, De Vooght, Vandierendonck, & D'Ydewalle, 2000).

It is worth mentioning that some logicians claim that the debate between advocates of mental models and mental logic has been carried out at the wrong level of abstraction. Stenning and Monaghan (2004) argue that mental models, mental logics and certain other systems for representing and manipulating propositions are all mutually translatable. Comparing the predictions for syllogistic reasoning of mental logic, mental models and a method of inference using Euler circles, they say that, "for every stage in using the representations in one method there is a comparable stage in each of the other methods." (Stenning & Monaghan, 2004, pp. 153–154) Agreeing with this, however, I do not agree with the conclusion they draw from it, that "This means that, in terms of the externally observed behaviour of people mentally solving syllogisms, it is impossible to say which method they are using." (Stenning & Monaghan, 2004, p. 154)

Stenning and Monaghan distinguish between the model theory level, at which the representations of mental logic, mental models and other systems are mutually translatable; a proof theory level; and the level of a theorem prover. I agree with this conceptual taxonomy: the vocabulary and syntax of a representation system are distinct from the rules that say what transitions between representations are allowed (the 'proof theory') and both are distinct from what strategies are used to determine which transitions to apply in which order in order to derive conclusions (the 'theorem prover'). Mental logic theory and mental model theory involve not just wellformedness rules for mental representations and lists of allowed transformations between them, but also strategies for forming the initial mental representations and for applying the translations. What prevents the theories from being notational variants is that they postulate different strategies, and that different transitions are claimed to be basic. One important sign of this divergence is the property of mental model theory mentioned above that it produces certain unsound inferences.

2.3.2 SOUNDNESS AND COMPLETENESS

As Braine has said (1978), mental logic theory draws on Gentzen's work in logic, which shifted the emphasis in logic from axioms to inference schemas, that is, from logic as a system built on a collection of foundational propositions, to a system which preserves truth. Gentzen was concerned that the inference schemas be psychologically real, reflecting "as accurately as possible the actual logical reasoning involved in mathematical proofs" (Gentzen, 1964, p. 291). According to Braine, the significance of Gentzen's work for the psychology of reasoning was not realised until decades later and was then the seed for mental logic:

> Gentzen's work went essentially unnoticed in psychological studies on reasoning until quite recently when a number of psychologists [Braine cites Johnson-Laird, 1975; Osherson, 1975b; Osherson, 1975a] have independently come upon it in the course of developing the essentially similar concept that proofs and chains of reasoning by human beings consist in the serial application of inference rules, and thus that a logical model for deduction should consist of a set of inference rule schemata (Braine, 1978, p. 3).

Both mental logic and mental model theory are concerned to show how in reasoning local consistency can be preserved while conclusions are generated. The solution that they give is that propositions are derived from other propositions in ways that resemble the rules of classical logic, either – in the case of mental logic – the syntactic rules of natural deduction, or, – for mental model theory – the semantic rules of truth-table proofs.

There are various dimensions on which such systems can vary. Two of the important ones are soundness and completeness. The transitions allowed by a system can be sound – truth-preserving – or merely heuristic. A system can be complete in that all logical entailments of a set of premises are deducible in the system, or incomplete, so that some logical entailments are not derivable using the rules available.

Mental logic is truth-preserving *ex hypothesi* since all transitions are governed by valid rules of inference. A mixed system is possible, however, with the addition to a mental logic of rules that are *not* truth-preserving. These heuristic rules might be domain-specific and sensitive to the content of the mental representations they operate on.

Mental models (to Johnson-Laird's surprise (1997a, p. 431)) sometimes generate false conclusions from true premises. (Johnson-Laird & Savary, 1996; Johnson-Laird, 1997b; Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000; Johnson-Laird, 2004, pp. 179–181, 191. See also discussion in Rips, 1997, pp. 416–417 and the reply, Johnson-Laird, 1997a.) The examples that have been discussed in the literature involve disjunctions and conditionals, where the non-representation of what is false leads to models from which false conclusions can be drawn.

(9) There is a pen or a book on the table, or else a book and a cup on the table.

There is a book and a cup on the table.

Is it possible that both assertions could be true at the same time? (this example and discussion are adapted from Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000)

The mental models of the first premise, the exclusive disjunction, are as follows:

Table 3: Mental models for exclusive disjunction of a disjunction and a conjunction

pen	
	book
pen	book
book	cup

The second assertion has one mental model, which is the same as the last line above. Therefore mental model theory predicts that participants should answer that the two assertions are compatible. The mistake is made at the stage of formation of mental models for the first assertion. The models formed do not contain as much information as the propositions they represent.

To see that the two assertions are incompatible, assume that the first clause of the first premise is true. In that case the second clause is false (taking 'or else' as an exclusive disjunction), so the second assertion, which is identical to the second clause of the first assertion, is false. Now suppose instead that the first clause of the first premise is false. In that case, there is neither a pen nor a book on the table, so the second assertion is false. So whether the first clause of the first assertion is true or false, the second assertion must be false: the two assertions must be incompatible, therefore.

Mental model theory also predicts illusory inferences with some disjunctions (inclusive or exclusive) containing embedded conditionals, as in the following example (based on an example in Johnson-Laird, Girotto, & Legrenzi, 2003):

(10) If there is a pen on the table then there is a book on the table, or if there isn't a pen on the table then there is a book on the table.

There is a pen on the table.

What, if anything, follows?

The prediction is that participants will say that it follows that there is a book on the table. An informal version of the procedure for working out what follows from these assertions according to mental model theory is as follows: the first assertion is interpreted as meaning that there are two types of situation. These are situations in which there is a pen on the table and situations in which there is not a pen on the table, and in both these types of situations there is a book on the table. The second assertion then tells the reasoner that the actual situation is the first one. In this situation the book is on the table, so it is concluded that it follows from the two assertions that the book is on the table. This conclusion does not follow, however, because the first assertion is a disjunction, so that from its truth one cannot infer the truth of either disjunct. Thus it is not necessarily true that if there is a pen on the table then there is a book on the table, so the situation where there is a pen on the table and no book on the table is consistent with the two assertions. It cannot, therefore follow from the assertions that there is a book on the table.

I think that this example can also be dealt with in terms of utterance interpretation followed by the operation of a mental logic. The utterance of the first sentence conveys something like *There is a book on the table whether or not there is a pen on the table* – which is truth-conditionally equivalent to Q: *There is a book on the table* – rather than the tautology that the sentence literally expresses, $(P \rightarrow Q) \vee (\neg P \rightarrow Q)$. Tautologies are not informative, so assuming that the speaker aims to convey something relevant, uttering a sentence that is tautological in form must be presumed to be intended to convey something other than the tautology itself. In this particular case, another way of seeing the conveyed meaning is that the 'or' is narrowed to 'and': $(P \rightarrow Q) \,\&\, (\neg P \rightarrow Q) = Q$.

Examples of this sort show that the procedures proposed by mental model theory are unsound, that is, merely heuristic. In general terms, this unsoundness results from the loss or non-representation of certain information about which possibilities exclude which others when the mental models are formed.

Mental models are also lossy[68] when representing propositions which involve quantification, and more obviously so. The model constructed from the representation/parsing of a sentence with quantificational elements will have specific instantiations of possible configurations of entities rather than variables. For example, according to the theory, "all *x*s are equal to the sum of some *y* and some *z*" is first parsed to give *(All x)(some y)(some z)(x = y + z)*. A model is then constructed by iterative choice of arbitrary values for the variables, constrained so that they fit the formula, for example [8 6](1 6 4 2)(7 7 2 2 4 4). (This example is from O'Brien, 2004, p. 226). The model is a particular instantiation of the formula, and does not capture its full meaning, so the formula could not be reconstructed from the model. The square brackets around 8 and 6 do not symbolize universality; and the model does not include the information that it is only one among infinitely many that could be created from the formula.

Although the reasoning procedure postulated in mental model theory can lead to invalid conclusions, it *often* leads to valid conclusions, at least when all the possibilities are represented as mental models. Thus mental models with exhaustive generation and search are close to algorithmic. Mental models with a stopping rule that terminates generation of mental models early would

[68] The term *lossy* is from computing. A lossy process (e.g. a compression algorithm) is one that loses information irretrievably. The opposite is 'lossless'.

be a less accurate but more frugal heuristic. At the most extreme, the generation of only one model, as proposed by Johnson-Laird for spontaneous deduction, is a simple, fast, frugal, and rather inaccurate heuristic.

Conversely, although the rules of a mental logic are sound, the theory can also account for erroneous conclusions, so to refute the theory it is not enough to show that people do not always reason correctly. It is puzzling then that "claims against the existence of a mental logic typically consist merely of showing that judgements of research participants have failed to correspond to some feature or other of a standard logic" (O'Brien, 2004, p. 207)

There are several reasons why this kind of evidence is at best inconclusive. First, mental logic might differ from standard classical logic, so that some inferences that are unsound in classical logic are sound in mental logic. I do not pursue this line. Secondly, reasoners may make performance errors due to capacity limitations of the system, interference from other mental systems, or disruption by non-mental causes like blows to the head. There are other possible sources of error, some of which were discussed by Henle and adopted by Braine in his earliest work on mental logic:

> Henle argued that deductive 'error' is due – not to illogicality – but to premises being omitted or interpreted in an unintended way, to the introduction of outside knowledge as an additional premise, or to a failure to accept the logical task. Thus, logical principles govern the movement from one step to another in an argument, but the 'effective' premises (the ones actually used by the subject) may not be the ones that the problem-setter intended (Braine, 1978, p. 2).

I have already explained in this chapter how the operation of deductive rules can lead to contextual conclusions which are not entailed by the presented information when extra premises are introduced by the reasoner. In deductive reasoning experiments, the instructions typically specify that what is of interest is what logically valid conclusions can be drawn from the premises provided, so it is required that one keep to what follows from the premises. In everyday reasoning, reasoners use all available information, and the habit is probably hard to break, so one would expect participants in reasoning experiments to frequently reach conclusions that are not entailed by the premises given. There is evidence that "more reflective and engaged reasoners will be more likely to affirm the axioms that define normative reasoning" (Stanovich & West, 1998, p. 293) and to obtain normative responses on reasoning problems. One can distinguish between *natural* reasoning using the reasoner's full range of knowledge, and *analytic* reasoning which uses only the special collection of attitudes, procedures and, perhaps, rules, that are applied to reasoning problems but not to everyday problems. (This terminology for the distinction comes from Braine, 1990).[69]

There is a second difference between the two modes of reasoning. Much depends on how the task is understood, in two distinct ways. The first consideration is how the information given is understood and mentally represented. In 'analytic' reasoning the premises must be interpreted as expressing the minimum possible commitment, so that, for example a sentence of the form 'if p, q' must be understood as a conditional, rather than a biconditional. In certain contexts, however, contextual factors may make it highly likely that the interpretation reached is biconditional. A more complex case is example (10) above. I have suggested that the 'illusory inference' generated in processing example (10) is due to pragmatic enrichment followed by the operation of deductive rules. This kind of consideration means that normative performance on reasoning tasks often depends on the ability to disregard implicatures or other pragmatically derived material that would normally be intended. A second interpretive factor is that the way the task is explained and set out will suggest how it should be attempted (which might include taking a non-deductive approach). I comment further in the next chapter on the way that pragmatic factors affect the tasks that participants attempt in reasoning experiments.

We have seen that systems that are logically sound, as well as systems that are logically unsound, are compatible with the observation that people often reach conclusions unsupported by the premises presented. Another difference between systems for inference is whether they are complete (in the logical sense). An incomplete system cannot itself generate all logical entailments of a

[69] Braine originally (1978) referred to these two modes as practical and formal reasoning. This use of the term 'practical' is not the one usual in philosophy, for reasoning about what is to be done (see section 2.2.6 above).

set of premises. Here mental logics vary. Some of the inference rules that would have to be included for completeness are implausible as psychologically basic rules. For example, most people are reluctant to endorse as valid an inference from a proposition p to a disjunction of that proposition with an arbitrary second proposition, $p \vee q$ (Rips, 1983; see also discussion of this point in Sperber & Wilson, 1986, pp. 99–100). The inference is valid but apparently not psychologically basic.

Rips' PSYCOP (Psychology of Proof) system is incomplete (Rips, 1997, pp. 418–419). It could be made complete by the addition of the inference schema in (11). Rips has said that this schema should not be considered part of the deductive system because it is not intuitively obvious (Rips, 1994; see also the discussion in Rips, 1997; Johnson-Laird, 1997b; Johnson-Laird, 1997a).

(11) $$\frac{\text{NOT (IF P THEN Q)}}{\text{P AND NOT Q}}$$

A second argument against the inclusion of a complete set of rules has been given by Sperber and Wilson (1986). Any complete set of rules includes some introduction rules, such as and-introduction (&I) and or-introduction ($\vee$I) (although not necessarily these particular rules). If we assume that the rules of a deductive system apply automatically to input of the correct form then introduction rules will lead to open-ended generation of trivial inferences[70] (on this point see also Johnson-Laird, 1997b, p. 392), rapidly overwhelming the computational resources of the deductive system, as in (12)[71]:

(12) a. $P \vdash_{\&I} P \& P \vdash_{\&I} P \& P \& P \vdash_{\&I} \ldots$

b. $P \vdash_{\vee I} P \vee Q \vdash_{\vee I} P \vee Q \vee R \vdash_{\vee I} \ldots$

Sperber and Wilson make the bold suggestion that the deductive system for spontaneous inference includes a mental logic with no introduction rules, where 'introduction rule' is defined as "a rule whose output contains every

[70] There may be some other tacit assumptions involved, as Uchida (2007) argues.

[71] This is a formal version of Harman's concern about trivial inferences, mentioned above.

concept contained in its input assumption(s), and at least one further concept." (Sperber & Wilson, 1986, p. 96) Then the computational explosion cannot occur.

Three other ways of dealing with this difficulty have been proposed. Johnson-Laird (1975) and Braine and O'Brien (O'Brien, 2004, p. 210ff) suggest systems in which some rules can only operate if they feed core rules, that is, if their output would be in the correct form for the core rules to operate on. In Braine and O'Brien's system, the core rules, which operate automatically on any representations in working memory with the right form, include modus ponens, double-negation elimination and inference rules that eliminate disjunctions and conditionals; the feeder rules for propositional logic are and-introduction and and-elimination.[72] In Rips's PSYCOP system, there are backward inference rules in addition to forward inference rules. The problematic inference rules are confined to backwards inference chains. Both of these proposals solve the problem without banning and-introduction, but at the cost of needing a substantive extra assumption in the theory, dividing inference rules into two or more classes, only one of which applies automatically to the contents of working memory. Braine earlier (1978) adopted a set of inference schemas (based on Gentzen's schemas) without and-introduction or or-introduction, but with other schemas that replicate their effects.

Barwise (1993) suggested that theories of reasoning must be logically complete. In a similar vein, Uchida (2007) argues that the inference system for pragmatics should be 'fully deductive', and that consistency requires completeness as well as soundness. Johnson-Laird's mental model theory is complete. Braine's original system is complete (1978, pp. 16–18), as is Braine and O'Brien's later system of mental logic (Braine, Reiser, & Rumain, 1984; Braine & O'Brien, 1998); Rips' PSYCOP system is nearly complete, as noted. Sperber and

[72] The division is into 'core schemas' and 'principal feeder schemas'. (There are also 'incompatibility schemas' and 'supposition schemas'.)

Both &I and &E are principal feeder schemas. These "are restricted so that they are applied only when their output provides for the application of a core schema." (p. 216)

In contrast, core schemas are "applied most freely by the reasoning program. [They are applied] so long as the propositions required for their application are conjointly considered in working memory."(ibid.)

The core schemas include a) DN elimination, b) if p_1 or p_2 or ... p_n then q; p_i therefore q, c) p_1 or p_2 or ... p_n, not p_i, therefore p_1 or p_2 or $p_{(i-1)}$ or $p_{(i+1)}$... p_n, d) MPP and others.

Wilson's deductive rules for spontaneous inference are less complete, in the intuitive sense that there are more entailments that the system cannot derive.

Two further considerations play a role in the theorist's decision to postulate a complete or incomplete system. One is whether the system is seen as defining the semantics of the logical connectives. A second consideration is whether the system is intended to account for all reasoning, or some limited subset of reasoning.

It is not necessary to take either mental logic or mental models as providing a kind of procedural semantics for logical operators in natural language, *pace* O'Brien's "the basic assumption of mental logic theory [is] that the meaning of a logic term is provided by its inference procedures" (O'Brien, 2004, p. 231). There is a clear conceptual distinction between the syntax of a logical language and an inventory of psychologically real rules that are logical in the sense that they preserve truth. All that is necessary for soundness is that each rule or schema in the mental logic is consistent with the semantics of the logical operators involved. The operators' semantics can be defined separately. There are some deep philosophical issues involved which I do not attempt to cover fully here[73], but it is worth saying that assuming that the semantics of logical operators is defined by the role they play in logical inference comes uncomfortably close to the psychologism that Frege opposed.

A further dimension of variation among theories of reasoning is the domain that they attempt to cover. Is the theory an attempt at an account of all kinds of reasoning, or at least, all types of theoretical reasoning? Or is the domain restricted to conscious effortful reasoning, or again to reasoning that is purely deductive? The area that interests me primarily, like Sperber and Wilson, is the spontaneous processing of information according to deductive rules, including abductive inference, since this is the domain of most utterance interpretation.

I agree with Braine that what is of primary interest is untrained reasoning, or at least the largely innate capacity for reasoning that all developmentally normal adults share. This clearly includes some facility to draw inferences that depend on logical connectives: "it is obvious that logically untrained subjects

[73] See Horsey, 2006, for thorough discussion of the relation between inferential and externalist accounts of the semantics of the connectives.

are able to reason with the English connectives, and it is this competence that a natural logic must capture." (Braine, 1978, p. 4).

The domain of reasoning in general (even restricted to theoretical rather than practical reasoning) seems too large and varied to be encapsulated in one set of rules, particularly given that there is good evidence that strategies can be learned and consciously applied. One can change how one reasons consciously by some combination of the following factors: applying learned strategies and shortcuts; effort of will; and using pencil and paper rather than doing it all in the head. As one would expect, research in psychology of reasoning shows that some of these factors change the results of reasoning. Experiments with protocols that allow participants to use pencil and paper or to report on their reasoning at the time may probe learned reasoning rules and strategies more than they probe the workings of natural deductive abilities. I think therefore that it is particularly problematic to place the emphasis on the investigation of conscious, effortful reasoning (which some theorists see as the only type of reasoning: see chapter 4 for Recanati's espousal of this view and its application to pragmatics).

2.3.3 CONCLUSIONS

Like Sperber and Wilson (1986) and O'Brien (2004), I tentatively adopt a mixed picture with a basic reliance on rules for reasoning isomorphic to the syntactic rules of logic, but with roles for heuristics that shortcut deductive rules, heuristics that direct reasoning, meaning postulates, and perhaps mental models. I do not think that any of the arguments that a mental logic must be complete are convincing, and I tentatively adopt Sperber and Wilson's assumption that introduction rules are not used in spontaneous inference. I assume that as well as rules of logical entailment, there are also meaning postulates: inference rules that allow the derivation of valid inferences based on conceptual content, such as the inference from *Mr Teeny is a monkey* to *Mr Teeny is an animal.* Indeed there is no reason not to see psychological deductive inference rules such as psychological *modus ponens*, and-elimination and so on as concept-based inference rules dependent on the logical information associated with the concepts of logical operators[74]. (On the parallel between logical and conceptual entailment see Sperber & Wilson, 1986, p. 84.)

I also assume that it is possible to adopt other strategies, including mental models, through direction of attention or learning. As Rips says:

> There is no doubt, for example, that people can learn devices like Euler circles or Venn diagrams and can use them to test syllogisms by searching for counterexamples. With practice, they can learn to manipulate these diagrams mentally, just as they can learn to do mental multiplication. (Rips, 1997, pp. 419–420)

Intuitively, mental models seem more likely to be used, perhaps as the basic mode of reasoning, in certain kinds of spatial problems. Empirical evidence for this thesis is mixed, however (Gilhooly, 2004, pp. 57–8, 66–71)[75].

Mental models might also be used, even in the absence of training, to check consistency when the mental logic used on its own gives no direct answer. In the absence of introduction rules, certain entailments will not be deducible. For example it will not be possible to deduce something of the form $P \,\&\, (Q \vee \neg Q)$ from something of the form P. Reasoners might nonetheless be able to work out that $P \,\&\, (Q \vee \neg Q)$ is true given that P is true by attempting and failing to find a model compatible with P but incompatible with $P \,\&\, (Q \vee \neg Q)$. O'Brien mentions some related uses for models:

> The ability to imagine a model that provides a counterexample to a supposition or to a possible inference made on extralogical grounds, for example, would be a valuable addition to one's reasoning skills. So also would be a strategy for proving the undecidability of a conclusion, which seeks two plausible alternatives that both are consistent with the premises, but with one being consistent with the conclusion and the other not. (O'Brien, 2004, p. 228)

[74] This amounts to unification of the aspects of rationality which Cohen (1992) refers to as *deductive* and *semantical.*

[75] Note that evidence that indicates that spatial representations are imagistic is not necessarily evidence in favour of mental models, which, while not propositions, are not images either.

These uses of mental models are procedural heuristics, rules of thumb that amount to discovery procedures. Similar rules of thumb might govern the use of deductive rules. One example is the procedure which Sperber and Wilson give for showing validity of an argument not derivable by deductive rules. To show that an argument is valid, it suffices to show that the premises are inconsistent with the negation of the conclusion. If the deductive device finds that there is an inconsistency between, for example, three propositions in working memory of the forms p, q and $\neg(p \& q)$, then it has established that the propositions of the forms p, q entail the one of the form $p \& q$ (Sperber & Wilson, 1986, p. 102).

In addition to heuristics that govern the use of deductive rules, I have suggested that there are heuristics that take their place. Such heuristics are effectively unsound inference rules which may be used to make inferences that, while unsound, are useful.

A theory that human reasoning competence includes mental logic is natural if one accepts that there is a logical format for mental representations, as O'Brien says (O'Brien, 2004, p. 206). I look at the case for a logical, symbolic representation format in the next section.

2.4 EXPLANATION AND MENTAL REPRESENTATION

2.4.1 REPRESENTATIONAL-COMPUTATIONAL THEORIES OF MIND

> ... if you admit that it's a matter of fact that some agents are rational to some degree, then you have to face the hard question of how they can be. (Fodor, 1985b)

> one true inference invariably suggests others (Conan Doyle, 1892a, p. 12)

In chapter 1, I endorsed the neo-Cartesian, Chomskian view of the mind/brain as a device that processes information from perception, generating mental representations. In perception these representations are of the object perceived. In the linguistic component of utterance interpretation the mind generates representations of the phrase uttered. In the current chapter I have endorsed a view of rationality as reasoning ability, and reasoning ability as primarily the ability to make truth- or warrant-preserving transitions. I have also looked at the mental logic research programme, which takes the view that reasoning involves the construction of chains of valid inferences as the basis for a psychologically and computationally realistic account of reasoning. In this section I focus on the thesis that central cognition relies on manipulation of representations in a structured symbol system. The domain of enquiry is broader than either deductive inference or spontaneous inference. Here the domain is central cognition as a whole, the area of thought behind rational behaviour and intelligent action.

There are three components to the hypothesis, all familiar from the discussion above on mental logic. The first is that there are mental representations with a propositional format, structured so that the form of each mental representation reflects the logical structure of the proposition it expresses. Thus the representations have predicate/argument structure and compound structure so that logical operators can be applied to constituents representing propositions and perhaps to sub-propositional constituents, as well. As O'Brien says, the mental logic programme shares with Fodor and with work in artificial intelligence:

> an epistemological assumption that in order for a declarative memory to exist, there must be a format for the storage of propositional information ... This format must be capable of representing properties and the entities that have these properties, and to keep track of which entities have which properties and which properties go with which entities. ... In other words, the mind must have some basic logical predicate/argument structure. Further, the mind should have some ways of representing alternatives among the properties and among the entities that have those properties, as well as conjunctions, suppositions, and negations both of properties and of entities ...(O'Brien, 2004, p. 206)

The second component of the idea is that intelligence is due to the manipulation of these mental representations according to their *formal* (or syntactic) properties, rather than their semantic ones. The first two assumptions are shared by much work in psychology, artificial intelligence and computer science, and are at the core of Fodor's Representational Theory of Mind (Fodor,

1975) and Newell and Simon's Physical Symbol System Hypothesis (Newell & Simon, 1976, p. 116ff; Simon, 1990, p. 3ff)[76]. The third component of the hypothesis is that although the processes manipulate mental representations according to their forms only, they can be such as to preserve the semantic value of the input representation. Fodor's version of the thesis is stronger than this: for him, the transitions between mental representations *must* preserve semantic value. There are two reasons for the weaker formulation that I have used. One is that much of cognition may be accomplished by heuristic shortcuts, as previously mentioned. I amplify on this below. A second reason is suggested by the work of Newell and Simon. For Fodor, firmly in the propositional (or logical) camp, reasoning, or intelligent thought, is the tokening of a series of propositional mental representations where the transitions between representations preserve truth, just as in mental logic theory.[77] Newell and Simon, concerned with problem-solving in a more general sense, do not necessarily require the symbol strings at each stage to have truth-values. One can see why by looking at the problems in computer science that they list as amenable to the Physical Symbol System approach:

> ... puzzles and games, operations research problems of scheduling and allocating resources, simple induction tasks..., chess..., systems that handle and understand natural language in a variety of ways, systems for interpreting visual scenes, systems for hand eye coordination, systems that design, systems that write computer programs, systems for speech understanding (Newell & Simon, 1976, p. 119).

Some of these problems are to do with perceptual rather than conceptual processing; some are arguably conceptual in a weak sense, but non-propositional, such as natural language parsing and chess-playing.

Since the focus of this thesis is the inferential component of utterance understanding, and I maintain that the input and output of this process must be

[76] Similar programmes have been given such names as the 'symbolic systems hypothesis' (by Rockwell), GOFAI (Good Old-Fashioned Artificial Intelligence) (by Haugeland) and 'High church computationalism' (by Dennett) (Rockwell, 2005)

[77] Fodor intends his hypothesis to cover what I would regard as non-inferential operations such as generation and transformation of phrase markers for sentences. (Fodor, 1987, pp. 144–145) I discuss his broad view of inference in chapter 5.

a conceptual representation or logical form (even if not always fully propositional)[78], it is possible to set aside non-conceptual processing. I will operate with a variant of Fodor's stronger hypothesis, generalized as discussed above so that it is value that is preserved, in line with Grice's suggestion that in practical and theoretical reasoning the aim is to preserve practical value and truth respectively (Grice, 2001, pp. 57–58), and Sperber and Wilson's point that manipulation of conceptual representations or logical forms should preserve warrant, in the same way that manipulation of propositions should preserve truth. Thus broadened, the hypothesis may be slightly narrower in its application than some theorists would prefer. As an example, consider a definition given by Gilhooly: "'reasoning' involves explicit sequential thought processes that are effectively equivalent to the application of a sequence of rules of some formal system" (Gilhooly, 2004, pp. 51–52), where formal systems include "deductive logic, mathematics, statistics, probability, decision theory... inductive and deontic logics" (Gilhooly, 2004, pp. 51–52). If my thesis were concerned with mathematical reasoning, then it might be better to broaden the hypothesis in this way. But it is important to keep sight of the points that 1) the rules applied in any of these formal systems are syntactic rules, that is, rules which operate on representations purely by virtue of the form of the representations, and 2) the rules generally respect semantic entailment, so that if the input is good (rationally acceptable), then the output is good (rationally acceptable) too.

The first of the three assumptions I have listed motivates the second and third: "Given the assumption that there is a logical representation format, one would also expect there to be some logical inferential processes" (O'Brien, 2004, p. 206) because intelligent creatures must be able to "make inferences that go beyond the presented information, and there ought to be some ways to ascertain which of these inferences are coherent." (O'Brien, 2004, p. 206) Logic serves the function of ensuring that "false propositions are not drawn from true premises" (O'Brien, 2004, p. 206) so one would expect that some of the rules for transitions preserve entailments[79].

[78] I attempt to justify the idea that the input to pragmatic processing is conceptual in chapter 4.

[79] See, however, Sperber's recent work (2000; 2001), in which he argues that reasoning

The key ideas are first, that "mental processes are causal sequences of mental states" (Fodor, 1985b, p. 91) and secondly, that the sequences are not simply causal, but that they share with logical arguments the property that each representation preserves the warrant of the ones that precede it. This means that trains of thought can be isomorphic to logical arguments (Fodor, 1985b, p. 91). As an example, Fodor cites a passage from the Sherlock Holmes story *The Speckled Band*:

> I instantly reconsidered my position when, however, it became clear to me that whatever danger threatened an occupant of the room could not come either from the window or the door. My attention was speedily drawn, as I have already remarked to you, to this ventilator, and to the bell-rope which hung down to the bed. The discovery that this was a dummy, and that the bed was clamped to the floor, instantly gave rise to the suspicion that the rope was there as a bridge for something passing through the hole and coming to the bed. The idea of a snake instantly occurred to me, and when I coupled it with my knowledge that the doctor was furnished with a supply of creatures from India, I felt that I was probably on the right track. (Conan Doyle, 1892b)

Here the thoughts that cause belief in a proposition are reasons for believing that proposition. Many theorists have thought that explanations for human behaviour cannot be causal (e.g. Winch, 1958; von Wright, 1971), in part because explanations are given in terms of reasons for that behaviour, and reasons and causes have different properties. First, to act as an explanation for behaviour, a reason must be understood by the agent. There is no such restriction on causal explanations. Secondly, reasons justify the behaviour that they cause; causes do not. What is at stake is whether the justificatory/explanatory role of reasons precludes them from also playing a causal role. Davidson's well-known argument that it does not (Davidson, 1963) is that a reason is not an explanation of an action, no matter how much it might justify the action, if it was not the operative reason. I might have a good reason for buying a fast, new computer: to get my thesis finished faster, for example. But if the actual

evolved under evolutionary pressure to assess the veracity of interlocutors, rather than to maintain coherence in the reasoner's own thought.

reason why I bought it was to play computer games, then the first reason is not an explanation of my action. The reason that explains the action is the one that was causally involved.[80] The symbol-system hypothesis can be seen as a way of incorporating this point into cognitive science. As Fodor says, "the syntactic theory of mental operations provides a reductive account of the intelligence of thought." (Fodor, 1985b, p. 98) Holmes' monologue is an example:

> What connects the causal-history aspect of Holmes' story with its plausible-inference aspect is precisely the parallelism between trains of thought and arguments: the thoughts that effect the fixation of the belief that P provide, often enough, good grounds for believing that P. (Fodor, 1985b, p. 92)

Holmes is engaging in 'reconstructive psychology' and his description of the train of thoughts amounts here to an argument for the conclusion reached. This distinguishes reasoning from another kind of train of thought, associative connections. As Fodor says, in an associative train of thought there is mental causation but not reasoning.[81]

Of course it is possible to reason about *someone else's* associative train of thoughts, as in a (rather fanciful) passage in one of Poe's detective stories (1841). Dupin, the detective, walks silently with a friend for some time, and then makes a comment on a subject that the friend has been silently considering[82]. Here the idea is that the two know each other so well that one of them can successfully infer what thoughts will be occasioned in the other by seeing a certain person and can also infer what mental associations will follow – and which will follow those, and so on, for several minutes. This example makes very clear the distinction between reasoning and a chain of thoughts driven

[80] See chapter 4 for more on Davidsonian causalism.

[81] In fact Fodor says something stronger: that associative sequences of mental representations are not *thinking*. I think that this use of the word defines thinking too narrowly – or we would not call an associative series of mental representations a train of thought.

[82] "'He is a very little fellow, that's true, and would do better for the *Théâtre des Variétés*.'

'There can be no doubt of that,' I replied unwittingly, and not at first observing (so much had I been absorbed in reflection) the extraordinary manner in which the speaker had chimed in with my meditations. In an instant afterward I recollected myself, and my astonishment was profound." (Poe, 1841) largely by associations. Here is part of Dupin's explanation of his own reasoning about his friend's thought processes:

> I knew that you could not say to yourself 'stereotomy' without being brought to think of atomies, and thus of the theories of Epicurus; and since, when we discussed this subject not very long ago, I mentioned to you how singularly, yet with how little notice, the vague guesses of that noble Greek had met with confirmation in the late nebular cosmogony, I felt that you could not avoid casting your eyes upward to the great nebula in Orion, and I certainly expected that you would do so. (Poe, 1841)

Here Dupin's own train of thought meets Fodor's criteria: each step follows logically from the previous one, given certain supplementary premises about his friend's knowledge of various subjects. Equally, the description implicitly presents each step as caused by the previous one. On the other hand, his friend's train of thoughts – at least if Dupin's description is accurate – is largely driven by associations. A thought about atomism gave rise to, indeed caused, a thought about stellar nebulae because of an association created, or reinforced, by a recent discussion.[83]

In stressing the difference between associative and logical trains of thought I do not want to suggest that any train of thought is in practice purely associative or purely logical. As discussed above, a great deal of inference involves the supplying of implicit premises, perhaps suggested by the context. Explanations for the storage and retrieval of this material postulate essentially associative links. Equally, any train of thought more structured than daydreaming will involve some inferential steps.

Nor do I want to say that no principles apply to both associative and inferential trains of thought. General principles of cognition will apply to both. For example, Sperber and Wilson propose that there is a cognitive principle of relevance: cognition is geared so as to tend to produce the greatest returns for the least effort by allocating resources to the contextual assumptions or im-

[83] It is true that cosmology does have something to do with particle physics (see, e.g., Collins, Martin, & Squires, 1989) but that does not mean that (the content of) the narrator's thought about atomism in any way implied (the content of) his thought about nebulae. (Of course the links Dupin and friend discussed in the mid 19th century are unlikely to be much like the links now understood to exist.)

plications that seem most relevant. If this is so, then both types of trains of thought will fall under that generalisation, the difference being that in a logical train of thought "the most relevant-seeming assumptions/implications happen to add up to a discursive argument", (Wilson, p.c.) whereas in an associative train of thought they do not.

2.4.2 FURTHER MERITS OF THE HYPOTHESIS

In this thesis, then, I adopt the RTM/symbol-system hypothesis on the basis that it provides a psychologically realistic account of cognition, including an explanation for the property of trains of thought in reasoning that there is a parallelism between the train of thought and a logical argument. The hypothesis has further advantages. It explains how information from the various senses, from memory and from utterance comprehension can be integrated. The proposal is simply that the information is all put into one format.

One traditional view is that natural language plays the role of integrating information. To the extent, though, that non-linguistic creatures such as non-human animals and pre-linguistic infants are able to reason, to make inferences, or to think intelligently, there is a need to explain how intelligent thought can occur without natural language. A structured symbol system for cognition, that is, a Language of Thought, is a useful explanation.

There are strong arguments against the idea that intelligent thought is literally conducted in natural language. It would be odd to import phonological (PF) features into reasoning and into other aspects of thought that are not subject to the constraints of the PF interface. Another consideration is that natural language sentences often underspecify the proposition they express. Indeed many theorists would say that they do not express any proposition. At the least, a natural language sentence would have to be disambiguated and have reference assigned to its indexical elements before it was suitable for use as a representation of fully propositional thoughts. The thought that a speaker expresses by uttering an ambiguous sentence is not itself ambiguous. The idea that we think in natural language, minus PF features, plus annotations marking disambiguations and reference assignment, is effectively a variation of the Language of Thought hypothesis. However it is not an especially plausible one, given the double dissociation between linguistic ability and general intelligence.[84]

It is one of the virtues of the standard Language of Thought hypothesis, where the Language of Thought is not a version of natural language, that it accounts elegantly for this double dissociation. If central cognition were mostly conducted in natural language (or a disambiguated, reference-assigned, unpronounced version of it) then one would expect linguistic impairment to pattern with general cognitive difficulties. In fact, there is a well-evidenced double dissociation between intelligent thought and linguistic ability. As Smith and Tsimpli write, "language can be impaired in someone of otherwise normal intelligence, and – more surprisingly – someone with intelligence impaired by brain damage may nonetheless have normal, or even enhanced linguistic ability" (Smith & Tsimpli, 1995, p. 3). Impairment of language together with normal intelligence is seen in Specific Language Impairment. Most autistic savants also have impaired linguistic ability, together with highly developed abilities in specific domains such as music, calendrical calculation, or drawing. Impaired intelligence together with normal or greater than normal linguistic ability is possessed by Williams syndrome children, 'chatterbox' children, Laura, and hyperlexics, "all of whom have great linguistic ability in the presence of severe cognitive deficits" (Smith & Tsimpli, 1995, p. 3). The subject of Smith and Tsimpli's study, Christopher, has poor general intelligence: he is unable to work out how to win at noughts and crosses and does not conserve number, but is highly gifted in the domain of language. His English is "essentially normal" (Tsimpli & Smith, 1998, p. 193). In addition he speaks, reads and writes twenty or more languages, several fluently. He acquires new languages rapidly, particularly their lexis and morphosyntax, with little practice.

[84] I am not committed to the idea that all concepts in the Language of Thought are innately specified, and I do not think there is any *a priori* reason why speakers of different languages should not have different concepts available to form mental representations with. Logical concepts such as conjunction, predication, negation and universality are presumably available to all, but they are presumably also expressible in all natural languages, so they need not be innately specified separately. It is also relevant that there are strong constraints on the meanings of newly coined words. The double dissociation evidence, however, suggests that the link between natural language and the Language of Thought is less direct.

2.4.3 REALISM AND EXPLANATORY POWER

Fodor and Newell and Simon both trace back to Turing and early computer science the history of the idea that cognition should be explained through computations over the syntactic properties of symbolic representations. The roots of the idea are in the formalization of logic of the late nineteenth and early twentieth centuries with its new stress on syntactic rules and formal symbol manipulation (Newell & Simon, 1976, p. 117). In various versions it emerged as the way of doing computer science and psychology in the nineteen-fifties and nineteen-sixties.

One can take the idea as an empirical hypothesis, as Newell and Simon do, or as a core assumption of a research programme[85], as it seems to me. Perhaps there is no great difference between the two views. Newell and Simon's examples from the history of science of generalisations with similar status to the Physical Symbol System suggest so, since in each case they are hypotheses that are at the core of research programmes: the germ theory of disease, the atomic hypothesis in chemistry, the cell doctrine in biology, and plate tectonics in geology. (Newell & Simon, 1976, p. 115) Newell and Simon describe these as "laws of qualitative structure." (1976, p. 115) Simon (Simon, 1990, p. 2) calls them "some of the most important invariants in science". They are qualitative rather than quantitative, and have many exceptions. Their function is to tell the scientist what type of explanation to look for:

> For example, the germ theory of disease, surely one of Pasteur's major contributions to biology, says only something like: "If you observe pathology, look for a microorganism that might be causing the symptoms." Similarly, modern molecular genetics stems from the approximately correct generalization that inheritance of traits is governed by the arrangement of long helical sequences of the four DNA nucleotides. (Simon, 1990, p. 2)

What Newell and Simon, and Fodor would agree on is that – in Fodor's terminology – the RTM/symbol system hypothesis is currently the only game in town.[86] If we wish to explain intelligence or rationality there is no well-de-

[85] In the sense of Lakatos (1970).

[86] Not everyone agrees, of course. See Rockwell, 2005 for a recent attempt to develop an alternative.

veloped alternative to a theory of manipulation of mental representations according to their syntactic properties.[87]

Explanatory power in psychology lies in having theories about the mechanisms that underlie behaviour, that is, having realist rather than instrumentalist theories. This criterion rules out such well-known alternative programmes as behaviourism and Dennett's 'intentional stance'.[88] No one ever showed how the theoretical apparatus of behaviourism could in principle account for intelligent thought. Furthermore, no one modelled, or even worked through, the details of any reasoning process in the behaviourist idiom (Newell & Simon, 1976).

Ryle's views are typical of the mid-twentieth century anti-mentalist tendency:

> Underlying all the other features of the operations executed by the intelligent reasoner there is the cardinal feature that he reasons logically, that is, that he avoids fallacies and produces valid proofs and inferences, pertinent to the case he is making. He observes the rules of logic, as well as those of style, forensic strategy, professional etiquette and the rest. But he probably observes the rules of logic without thinking about them." (Ryle, 1949, p. 48)

All of this is true, but does not support the implied conclusion that there is no need for an explanation of intelligent thought in terms of mental representations. Intuitively, it seems true that (for most reasoning) a reasoner (even an intelligent one!) observes the rules of logic without thinking about them, although sometimes a reasoner may think about, or even reason on the subject of the rules of logic. (And metalogicians reason about principles constraining the rules of logic.) But Ryle avoids the question which is important for a scientist: *why* does a reasoner obey the rules of logic? That is: what is it about humans beings which causes them to follow the rules of logic (when they do)?

There seem to be two answers possible in principle. Either (1) the rules of logic are known to the reasoner i.e. they are mentally represented. They might

[87] Fodor has provided arguments against connectionist alternatives (e.g. Fodor, 1987), which I agree with but do not discuss here.

[88] Newell and Simon make the same point about Gestalt psychology (Newell & Simon, 1976, p. 120).

or might not be consciously accessible: that is not what is at stake here. Chomsky's view that the principles of syntax are known (or 'cognised') is an example of a theory or research programme that postulates that mental activity according to certain principles is due to explicit (although not conscious) representation of the principles in the mind. Alternatively, (2) the rules of logic are not mentally represented but some properties of the reasoner's mind mean that when it works it follows the rules of logic. This second kind of explanation has also been given in the study of natural language syntax.[89] For example, van de Koot and Neeleman argue that:

> the grammar and the performance systems are theories of the same object, but at different levels of description: the cognitive and computational level, respectively. More precisely, the language faculty consists of encoding and decoding devices and the grammar is the code they adhere to. It can be shown that, if well-organized, the computational level is unlikely to contain a separate knowledge base. Rather, grammatical principles can be seen as emergent properties of natural language computations. (Neeleman & van de Koot, 2004, p. 1)

I am not committed either way on the status of the rules governing transitions in central cognition. The mental representations that the theory insists on are the thoughts in the sequences of thoughts that form valid arguments. The rules governing these transitions might be mentally represented, or they might be emergent properties of the reasoning system, only represented explicitly in our scientific theories of reasoning. Some of the rules, meaning postulates, for example, may be mentally represented and some not, perhaps

[89] Ryle, of course, also denied that mental rules are causally involved in linguistic activity, offering the consideration that "a foreign scholar might not know how to speak grammatical English as well as an English child, for all that he had mastered the theory of English grammar" (Ryle, 1949, p. 41). A similar passage is:

> I could not now read a Greek sentence, if I had not formerly learned Greek grammar, but I do not ordinarily have to remind myself of any rules of Greek grammar, before I construct a Greek sentence. I construe according to these rules, but I do not think of them. (Ryle, 1949, p. 315)

As with Ryle's comment in the text about rationality, all of this is true (at least if 'think of' means something like 'think about' and if mastering the 'theory of grammar' for a language is a matter of knowing a description that is not generative), but does not support Ryle's intended conclusion.

including the core logical rules. It might be an emergent property of our reasoning systems that we are disposed to reason from *P & Q* to *P*, but an explicitly represented rule that '*x* is a monkey' entails '*x* is an animal'. The rules must play a causal role, but they might do so in the way that instructions in a computer program do, or in the way that the law of gravity does.

The criterion that a causal explanation is sought militates against Dennett's 'intentional stance', as Fodor argues (Fodor, 1985b, pp. 79–81). For Dennett, talk about beliefs and desires is not to be taken as describing the internal structure of agents, but as adopting a stance towards them, treating them *as though* they had such mental states (Dennett, 1971; Dennett, 1987). The contention is that this provides a basis for understanding the behaviour of agents, including predicting how they will act, even though they do not actually have such states.

Dennett's instrumentalism about such mental representations as beliefs and desires seems to be based on a classical version of rationality (Cherniak, 1981, pp. 162–163). The idea is that to explain behaviour in terms of beliefs and desires we need to assume that the agent is fully rational (i.e. rational in a classical, unbounded sense). An unboundedly rational agent has a consistent set of beliefs, a consistent set of preferences, and acts to maximize his returns. Real agents are not unboundedly rational, as I discuss in the next chapter. On the other hand, Dennett thinks that agents' behaviour must be close enough to rationality for evolutionary reasons: thoroughly irrational creatures would have been selected out. So explanation in terms of beliefs and desires will be close enough for predictive purposes even though it is not actually true (Fodor, 1985b, pp. 79–80). Thus, Dennett says, we think about other people (and other agents) *as if* they had beliefs and desires. This might be a good explanation of what people do when explaining others' behaviour, but our interest is not in explaining people's explanations of intelligent behaviour, but in explaining the behaviour itself. To do that we cannot simply say that because of evolutionary pressure people's behaviour will be rational, or mostly rational: we have to attempt to explain the behaviour in terms of mental structures.

There is a marked similarity with Simon's criticism of 'as if' theories in economics and other social sciences, specifically decision theory and rational choice theory[90]. Rational choice theory assumes that humans choose the best action given their preferences and the choices available. Thus to know what an agent will do it is sufficient to know what it is best for him to do, given his preferences and the environment in which he is choosing. Simon argues, on the contrary, that what is necessary for theoretical understanding of intelligent behaviour are theories that attempt to describe the mental mechanisms responsible (Simon, 1990, p. 6ff). I discuss this line of argument in more detail in chapter 3, where I discuss optimising theories as a species of classical rationality.

2.4.4 FODOR'S PESSIMISM ABOUT CREATIVE, UNBOUNDED, CENTRAL THOUGHT

In this thesis, then, I assume that central cognition, including reasoning, should be thought of as involving a series of mental representations in which one representation (or set of representations) causes the next in virtue of its form and that in these chains of transitions, accomplished by purely syntactic means, semantic entailment is mostly preserved. As previously remarked (in section 2.2.2), though, the particular central system which this thesis attempts to understand is not a purely deductive reasoning system, since it performs non-demonstrative inferences. I have argued there that the picture of a series of mental representations linked by rules isomorphic to rules of derivation cannot be the whole story, but that it is a part of the story. Deductive steps, from the premises supplied taken together with contextual information, lead to contextual inferences.

It is because of the differences between analytic, deductive reasoning and synthetic, non-demonstrative reasoning that Fodor is pessimistic that central cognition can be understood in terms of his Representational Theory of Mind. He takes scientific theorising to be a paradigm central cognitive activity, and, as remarked above, an important part of scientific theorizing is coming up with hypotheses which explain and predict but (famously) do not logically follow from observational data. That is, a key part of scientific theorizing consists in abductive inference, inference to the best explanation. Some other

[90] There are some historical connections between rational choice theory and behaviourism. Homans, a pioneer in bringing rational choice theory to the social sciences beyond economics, espoused a behaviourist psychology (Scott, 2000, p. 127).

central processes appear to be similar in this respect. Mindreading involves postulating explanations for others' behaviour in ways that go beyond the data. If I see my flatmate heading for the fridge late at night, I may infer that he is hungry and looking for a midnight snack. My observation somehow triggers my forming a hypothesis about his behaviour, but the hypothesis is not entailed by the behaviour observed. That the same goes for utterance interpretation was an important part of Grice's message, as discussed in chapter 1. Any observations a hearer may have made of an utterance fall short of logically entailing the meaning of that utterance. So the speaker meaning which the hearer arrives at is a hypothesis and the process is inference to the best explanation.

Now abductive inference clearly involves a creative element over and above any creativity demanded by deductive inference. Deductive inference is somewhat creative. In such forms of deduction as proving logical sequents by inference rules, there is an element of creativity in that a choice must be made at each step as to which rule to apply.[91] In abductive inference the involvement of creativity is qualitatively much greater. It is not just a matter of how to manipulate the information one starts with. Seemingly unrelated information must be brought in. In the example given, my flatmate's mental representations of the contents of the fridge are invoked as part of the explanation for his behaviour. Scientific explanation provides many examples in which there was a considerable creative leap involved in hypothesising a causal link. In chemistry and in thermodynamics, for example, the properties of solids, liquids or gases are often explained in terms of statistical generalisations about *prima facie* unrelated properties of small objects that they are composed of. The creative leap involved is considerable given that the smaller objects had not been observed when the theories were formulated. Abductive explanations can be tested once they exist. But in order to come up with them, certain leaps must be made.[92] This kind of creativity is presumably at least part of what Fodor

[91] In propositional logic the process of deduction can be mechanised, but in other forms of logic, including first-order predicate calculus, there is no determinate procedure for proofs of sequents (Lemmon, 1978, p. 91; Gamut, 1990, p. 150f): a theorem prover may not terminate in a finite number of steps.

[92] Generally, problems that are undecidable, or too computationally expensive to solve by brute force methods (most interesting problems have both properties) must be solved by

means when he says "what is most characteristic, and most puzzling, about the higher cognitive mind[93] [is] its nonencapsulation, its creativity, its holism, and its passion for the analogical." (Fodor, 1985a, p. 4)

Central systems responsible for tasks such as mindreading and utterance interpretation take input from peripheral systems and reason from it, accessing memory, in order to provide explanations for the perceived phenomena. This process of generating explanations is non-demonstrative and highly dependent on the associative or analogical processes which generate candidate hypotheses (Fodor, 1983, p. 107). Fodor calls this aspect of cognition 'Quinean' and 'isotropic', where by 'isotropic' he means approximately[94] that in principle any information may be relevant to the outcome of a conceptual process (Fodor, 1983, p. 105) and by 'Quinean' he means that the criteria that are relevant to judging the goodness of a hypothesis are global properties such as the simplicity of one's belief system (Fodor, 1983, pp. 107–108).

Simon recognised that many thought that these areas provide the most serious challenge for a symbol theory of cognition, but thought that the problems had been solved in principle. The solution involves a second hypothesis – or 'law of qualitative structure' – about intelligent thought: that problem solving is a matter of heuristic search, some combination of trial and error. To find a solution that solves a problem, an intelligent system generates solutions and tests them, one by one. The problem of intelligent creativity is the problem of doing better than one would do generating solutions at random. How is it that, in certain domains, cognition is tuned to provide fruitful postulates? There are several parts to this solution of this problem, at least as it relates to many tasks, including utterance interpretation[95]. One is that the generator

some combination of trial and error:

> why not simply generate at once an expression that satisfies the test? This is, in fact, what we do when we wish and dream. "If wishes were horses, beggars might ride." But outside the world of dreams, it isn't possible. To know how we would test something, once constructed, does not mean that we know how to construct it – that we have any generator for doing so. (Newell & Simon, 1976, p. 121)

[93] By the higher cognitive mind, Fodor means those mental faculties (or that mental faculty) that deal(s) with conceptual rather than perceptual mental phenomena, i.e. central cognition.

[94] Fodor writes, "It is notoriously hard to give anything approaching a rigorous account of what being isotropic and Quinean amounts to." (1983, p. 105)

[95] Another problem for which the first two elements are important is choosing a chess move.

effectively has built into it some of the tests that the solution should satisfy. In this way, only solutions that will pass the tests are generated: there is trial, but not much error. In the case of utterance interpretation, one such incorporated test is that the implicated premise(s) and explicit meaning of an utterance together logically support the implicatures. This property is built in to the generator because it makes use of deductive rules to generate implications. The second part of the solution is what Simon calls 'recognition'. The idea is that reasoners store rich and extensive data about the problem domain, and that knowledge of patterns can be substituted for search.

The question of creativity can be seen as the question of how to reduce the portion of the problem space that is actually searched. The solution may involve canonical deductive rules guided by heuristics, and sometimes shortcut by heuristics. The process must rapidly generate and evaluate solutions. The starting point of search may be close to the solution because the first trial solution is fed by domain-sensitive recognition of patterns. The way to see if a model of this sort works is to take a particular area of reasoning, e.g. inference to the best explanation for utterances, and see if it can work there, as I do in chapters 4 and 5 below.

In this last section of this chapter I have set out one law of qualitative structure for cognitive science: the RTM/symbol system hypothesis. Towards the end, I have sketched a way to answer some of Fodor's scepticism about explaining central cognition in these terms by adopting a second law of qualitative structure, heuristic search, which I return to in more detail towards the end of the next chapter. There is no reason to think that in practice, quick, automatic abductive reasoning is Quinean or isotropic. In principle, any information might be relevant, but in practice, for certain tasks at least, cognition is tuned so that only information that is likely to be relevant is used. Again, in principle, one might judge the goodness of a solution by its coherence with one's entire belief system, but in practice the criteria are more local, defined by the task.

Good players rely on recognition, storing perhaps 50,000 distinct patterns (Newell & Simon, 1976, p. 125), and only potentially good moves (and only legal moves) are considered.

Chapter 3 · Classical and bounded rationality

> The question is how you arrive at your opinions and not what your opinions are. (Russell, 1983, p. 91)

In the introductory chapter I claimed that utterance interpretation is a boundedly rational process. Bounded rationality is a tendency in theorising about rationality which recognises the fact that "humans are in the finitary predicament of having a fixed limit on their cognitive capacities and the time available to them" (Cherniak, 1981, p. 165) and stands in opposition to another tendency: classical rationality. As mentioned above, advocates of bounded rationality try to explain judgements and choices in terms of heuristics: procedures which amount to shortcuts. They also stress the finding of solutions which are good enough, rather than optimal, i.e. that satisfice (in a broad sense). These two aspects of the programme are separable: not all heuristics find solutions that are 'good enough' – some do not do well enough, and some may overachieve relatively, finding optimal solutions with minimal search. In addition, logically one could satisfice by thoroughly examining all alternatives and picking one that is good enough but sub-optimal, so satisficing does not *require* heuristic shortcuts. What is more, some heuristics are fast and frugal and others may be lengthy and costly.

The key idea of a programme of bounded rationality is that not all of the problem space is searched. Problems are solved by generation and assessment of trial solutions, typically sequentially. Frugal heuristics make use of recognition of the type of situation to limit the number of solutions tried. Some such heuristics may approach the limit of frugality, at which the first solution generated will usually be chosen. In section 3.3 I look at several classes of boundedly rational procedure along the lines set out here.

However, before moving on to examination of types of boundedly rational procedure, I discuss some reasons for adopting the bounded rationality programme in the first place, looking at the competition, classical theories of rationality. The programme of bounded rationality is a reaction (originally by Simon 1947; 1969; 1982; 1983; 1990) to previous work which presents as a received view an idealised picture of rationality in philosophy and, particularly, in economics.

It is convenient to refer to this latter type of model of rationality as 'classical'. There is some risk of inaccuracy in speaking this way, since philosophers and economists have tended to idealise rationality in somewhat different ways, and, as one would expect, within each of these broad disciplines there have been different views of rationality. There are common elements however, across and within disciplines, which make classical visions of rationality similar to each other and distinct from bounded rationality.

Classical visions of rationality assume that a rational agent has consistent beliefs and consistent preferences, and finds solutions that are both logically or probabilistically normative and also optimal. This view faces difficult theoretical and empirical questions. The empirical evidence has been taken to show that humans do not have even basic logical competence. In section 3.2, I argue that this bleak view is unjustified. The evidence is that we are capable of good reasoning, within the limits one would expect of finite creatures. In the first section of this chapter, I look at theoretical arguments which also suggest a bounded view of rationality.

3.1 THEORETICAL CONSIDERATIONS

The classical vision of rationality emphasises consistency and optimisation, in contrast to the focus of theorists of bounded rationality on simple procedures and satisficing. Consistency is a property of a system, for example the agent's belief system, or his system of preferences, or his beliefs and intentions taken together as a system. The ideal of consistency is context- and content-independent, applying across domains. A classically rational agent is one who satisfies consistency conditions on beliefs and preferences such as the following: do not believe propositions *p* and *not-p*; do not simultaneously prefer outcome *a* to outcome *b*, outcome *b* to outcome *c*, and outcome *c* to outcome *a*; do not rate the conjunction of two events, *x* and *y*, as more likely than either one occurring.

Optimisation, on the other hand, is a constraint on aims or outcomes. A requirement to optimize is a requirement to find the best solution to a problem, to make the best choices, and generally to do as well as it is possible to do. Optimisation is not independent of context and content, since the best judgement will always depend on what is available. However it is generally insensitive to context and to the specific content of the problem. To be sure that the best outcome is reached it is necessary – in the general case – to weigh up all possible outcomes, in the light of all information that might be relevant. Thus a classical optimizer would generally have to carry out exhaustive search, regardless of the context.

There are links between these two pillars of classical rationality. One connection is that the requirement that an agent's beliefs are all consistent with each other is a requirement for a form of optimisation. As we shall see, meeting this requirement is computationally impractical, so it is an unrealistic criterion for rationality.

A second link between optimisation and consistency is fundamental to decision theory. Decision theory (and much of economics, and related work in other social sciences[96]) views rational agents as those which have consistent preferences and maximize their utility (or their expected utility). In a widely prevalent interpretation of decision theory, maximisation (of expected utility) follows from internal consistency of preferences as long as the agent chooses what he prefers. Here maximisation of expected utility is an optimization: rational agents are supposed to make the best choices, that is, those that bring them the greatest possible returns. This requirement to optimize need not be stated as an axiom of rationality, however. Rather it emerges from the requirement that an agent's preferences are consistent in a certain way. "On certain decision theoretic approaches... rationality requires only that one's preferences meet certain ordering criteria" (Mele & Rawling, 2004a, p. 4). Preferences that meet these criteria automatically maximize, as long as the agent acts according to his preferences, i.e. chooses what he prefers.

[96] The field of work in which the framework of decision theory is applied to (e.g.) sociology and political science is called 'rational choice theory'. See Scott, 2000.

It is not immediately clear whether this theory is descriptive or normative.[97] Is the theory a description of the behaviour of people, or a standard that people should aim at? Saying that 'a rational agent' behaves in a certain way allows for either interpretation, or some blend of the two. Daniel Ellsberg, who showed that not all uncertainty can be reduced to (quantifiable) risk, summarised the consensus against which he was arguing:

> The propounders of these axioms tend to be hopeful that the rules will be commonly satisfied, at least roughly and most of the time, because they regard these postulates as normative maxims, widely-acceptable principles of rational behavior. In other words, people should tend to behave in the postulated fashion, because that is the way they would want to behave. At the least, these axioms are believed to predict certain choices that people will make when they take plenty of time to reflect over their decision, in the light of the postulates. (Ellsberg, 1961, pp. 645–646)

Since the work of Kahneman and Tversky, however, it has been generally accepted that human behaviour deviates in certain ways from the classical picture. For example, people are generally risk averse and prefer to reduce risk even at the cost of lowering expected returns (Kahneman & Tversky, 1979).[98]

The tendency in philosophy has been to see classical rationality as primarily normative. According to this way of thinking about rationality, rationality is largely a matter of the conformity, or otherwise, of one's beliefs, desires, intentions and other mental attitudes to certain standards. Beliefs should be justified and consistent with each other; intentions should be compatible with one's beliefs and desires, and one should try to maximize their fulfilment.[99]

[97] This is a general problem for economics. Thus, for example, Hausman (2006), places "Positive versus normative economics" at the head of his list of methodological problems faced by economics.

[98] Ellsberg had earlier demonstrated another deviation from maximisation of expected utility, ambiguity aversion (Ellsberg, 1961). The paradox Ellsberg uses to demonstrate this was known to Keynes (Keynes, 1921, pp. 75–76, 315 fn 2).

[99] Consistency requirements are sometimes presented as conceptual necessities. For example, Davidson gives a rather Quinean argument in support of transitivity of preferences:

> I do not think that we can clearly say what should convince us that a man at a given time (or without a change of mind) preferred *a* to *b*, *b* to *c* and *c* to *a*. The reason for our difficulty is that we cannot make good sense of an attribution of preference except against a background of coherent attitudes. (Davidson, 1980b, p. 237)

Agents are rational to the extent that they meet these standards. It is compatible with this view that in reality most agents fall short in one way or another from time to time[100], although it is a fairly recent development to stress as Cherniak does that real humans cannot meet these idealised standards:

> Until recently, philosophy has uncritically accepted highly idealised conceptions of rationality, But cognition, computation, and information have costs; they do not just subsist in some immaterial effluvium (Cherniak, 1986, p. 3). ... the pervasively and tacitly assumed conception of rationality in philosophy is so idealized that it cannot apply in an interesting way to actual human beings (Cherniak, 1986, p. 5).

A further difference with classical rationality in economics is that decision theory is generally concerned with consistency and optimisation as they apply to preferences and choices rather than to systems of belief, whereas classical rationality as set out by philosophers concerns both.

Regardless of differences of this sort, a strict division between philosophers' and economists' conceptions of classical rationality would be artificial. Philosophers have made substantial contributions to debates in choice theory (in particular, Nozick, 1973; Nozick, 1974), and economics has always drawn on philosophical conceptions of rationality. Indeed much of economics and (more recently) game theory can be seen as detailed attempts to answer "an old hypothetical question" (Sen, 1977, p. 319) debated since at least the eighteenth century by philosophers and theologians as well as economists, "namely, in what sense and to what extent would egoistic behavior achieve general good?" (Sen, 1977, p. 321).

Against such Quinean arguments, which are also made about the attribution of inconsistent beliefs, it is worth noting that correct interpretation of a person's utterances may (*pace* Quine) attribute inconsistencies (and falsehoods) to him (Cherniak, 1986, p. 56). There are examples (noted by Sperber & Wilson, 1986, pp. 197–200) in which an implicated premise is needed to make sense of the utterance. The implicated premise may be false or thought by the hearer to be false, inconsistent with some of the speaker's other beliefs or thought by the hearer to be inconsistent with some of the speaker's other beliefs.

[100] It is often said that there is a minimum standard, as well. More as a matter of definition than description, a system or being will not count as an agent if its beliefs and actions do not meet minimal standards of consistency.

In recent decades, ideas about rationality have tended to flow from decision theory, economics and game theory to philosophy and to social sciences. The common perception is that substantial progress has been made in those fields with the assumption of a particular view of rationality; and this view of rationality has become influential outside these fields as a result. In particular there has been widespread interest in what game theory says about interacting agents and its explanation of the way that individually rational behaviour can lead to socially sub-optimal outcomes. In a recent philosophical survey of rationality (Mele & Rawling, 2004b), almost a third of the papers, seven of the twenty-two, discuss decision theory, economics or game theory. There are assumptions in the air that if a situation involves interactions between agents, then a game-theoretic treatment is a natural move, and in any situation where the preferences of individuals are to be investigated the axioms of decision theory should apply.

Thus in a discussion of the current state of the classical view of rationality, even as it bears on utterance interpretation, it is important to give some space to the decision-theoretic view of rationality. Even though decision theory is more concerned with preferences and actions than beliefs, and utterance interpretation is an exercise of theoretical rationality, many would assume that a game-theoretic treatment of utterance interpretation is natural. I have given specific arguments against such a treatment elsewhere (Allott, 2006). Here I look at problems with the classical, idealised view of rationality at the root of these views.

A further reason for looking at decision theory is that it renews an old challenge to psychological realism as a methodological commitment. Classical conceptions of rationality sit more easily with a methodology that is agnostic about mental representation than bounded rationality does. Bounded rationality stresses the processes involved in reasoning and decision making, while for classical rationality what matters most is that decisions and judgements are optimal and consistent: how they are reached may be abstracted away from. Decision theory and game theory make this aspect of the classical vision very clear. As I explain below, decision theory is methodologically agnostic about the mental representations behind choices, stressing instead the formal properties of the preference relation. It hardly needs saying that it is more of a live research programme than behaviourist psychology. It is controversial here, as it is not in psychology, to argue that the form of mental representation and the procedures used in reaching judgements cannot be ignored.

In the end, I argue, advocates of a bounded view of rationality have convincingly shown that classical visions of rationality are not descriptively correct. People could not and do not possess or maintain completely consistent systems of beliefs, intentions or preferences. For quick, everyday decisions, people could not and do not optimize in the classical sense, acting as though they actively considered (or considered whether to consider) all possible solutions to a problem and all potentially relevant information. However, it is not so easy to show that classical rationality is not the normative standard for rational agents: even if no one is classically rational, classical rationality might still have a normative force. There would be something strange, however, about a norm that no one met, or could meet. If we adopt a vision of rationality as bounded, there will be consequences for our view of rationality as a standard for people's reasoning and behaviour.[101]

Instead of optimization, advocates of bounded rationality stress economical processes which reach answers that are good enough, where what is good enough depends on the task and the context. The emphasis is on how properties of the process enable it to reach good answers for a given task rather than on formal properties of the system or the outcome. Instead of seeing consistency as the key property of systems of belief or preferences, they see it as, at most, one among other properties. For example, Gigerenzer argues that consistency is at most only a secondary criterion for good decision making, coming well behind "accuracy, speed, frugality, cost, transparency, and justifiability" (Gigerenzer, 2001, p. 3007).

The positive programme of bounded rationality set out by Simon, and considerably advanced recently in the work of Gigerenzer and colleagues, aims to show how rational decisions can come out of psychologically plausible mechanisms, where to be psychologically plausible a mechanism must be

[101] Gigerenzer writes, "even critics have generally retained the beautifully simple principles drawn from logic and probability theory as normative, albeit not descriptively valid – that is as definitions of how we should reason." (2000, p. 202) Gigerenzer thinks that they should be given up as norms as well.

computationally tractable. Generally consistency and maximization will only be local properties, if they are present at all, because global checking and global search are computationally intractable, and therefore psychologically implausible.

In the next section I look at some internal criticisms of decision-theoretic version of classical rationality, including the views of Amartya Sen. Although Sen is not an advocate of the positive programme of bounded rationality – satisficing and a focus on procedures – his work is is akin to bounded rationality in that he rejects consistency and optimization as the central pillars of a theory of choice.[102] In particular, Sen criticises the idea that agents must have consistent preferences, regardless of context.

3.1.1 DECISION THEORY AND GLOBAL CONSISTENCY

> Leibniz's dream was of a formal calculus of reasonableness that could be applied to everything. Modern variants tend to go one step further and assume that the calculus of rationality has already been found and can be imposed in all contexts. (Gigerenzer, 2000, p. 2002)

The classical tendency in economics assumes that rational agents obey the laws of logic and probability (Gigerenzer, 2000, p. vii), so, for example, agents have transitive desires and their choices are internally consistent (Gigerenzer, 2000, p. 202). Economists place less stress than some philosophers do on the external justification of desires: a rational agent in economics is one whose preferences are internally consistent and who acts so as to fulfil his or her desires to the greatest degree possible, whatever those desires may be. Such a position is often called Humeanism or Humean instrumentalism by philosophers, because of remarks in Hume such as "reason is and ought only to be the slave of the passions" (Hume, 2003, p. 295). The idea is that one can, given goals or desires, reason about how best to fulfil them, but that one cannot by reasoning alone reach any decision about what one's goals or desires should be. This is, of course, a controversial position; and to be a decision theorist one does not have to agree with it. Decision theorists are *methodological*

[102] Sen's positive programme is related. He wants economics to investigate how properties of choice and preference are driven by aims, beliefs etc.

Humeans in that they look only at the conformity of the preference relation to axioms, not how the agent arrived at that preference relation (nor how it is mentally represented or mentally processed).

For most economists and decision theorists, it is maximisation of (expected) utility that is the hallmark of a rational agent: it is assumed that rational agents make choices which are maximally fulfilling of the desires that they have, whatever those happen to be. This is because, for a decision theorist, a rational agent is an agent who makes sure that her preferences conform to the axioms of choice theory: "Insofar as decision theory has any normative judgments to make, any advice to give, it is best to think of it as telling us to conform our preferences to its axioms." (Dreier, 2004, p. 160) These include the requirements that the ordering of preferences is complete, and that the preference relation, R, which holds between any two alternatives, x and y – so xRy means 'y is not preferred to x'[103] – is antisymmetric, reflexive and transitive. For any agent for whom all the axioms hold, that agent's preferences are consistent and coherent, and there is a family of expected utility functions which express those preferences. So the axioms provide a standard for rationality as far as preferences are concerned, and this standard stresses coherence[104]. According to this picture of rationality, "[a rational agent] never has to *try* to maximize her expected utility. If her preferences conform to the axioms, then the maximisation of her utility will take care of itself (as long as she chooses what she prefers!)" (Dreier, 2004, p. 160).

There are methodological and formal advantages to this view of rational agents. Methodologically, one can find out what an agent values by seeing what he or she maximizes, subject to the assumptions that it is expected value of one sort or other that is maximized, and that the axioms of decision theory hold.[105] If an agent consistently chooses bananas rather than apples, or in-

[103] i.e. either the agent prefers x to y or is indifferent between them: as far as that agent is concerned, the desirability of x is greater than or equal to the desirability of y.

[104] Not every economist who accepts the axioms sees them as criteria for internal consistency of choice. They can be seen as following from the requirement that expected utility be maximized rather than the other way around. See Sen's remarks on the pioneer of revealed preference theory, Samuelson (Sen, 1993, p. 497, fn 5).

[105] The methodological advantage is obtainable only at the expense of a certain simple-mindedness theoretically, as Sen points out:

> If you are observed to choose x, rejecting y, you are declared to have 'revealed' a prefer-

creased leisure rather than longer working hours with more pay, then we can infer the agent's preferences without any need to ask him what he prefers. The formal advantages include the possibility of proving certain results about an agent who conforms to the axioms: one such result is the consequence mentioned above, that the agent's preferences are expressed by an expected utility function. Another way of seeing this advantage is that decision-theoretic rational agents are 'known quantities'. Therefore one can show (indeed prove) what they will do if they interact, as in mainstream economics and game theory.

Criticisms

There have been powerful criticisms of the decision-theoretic vision of rationality. Some of the criticisms, such as Simon's attack on optimising and 'as if' theories, challenge the plausibility of this picture of rationality as a whole. As I have indicated, I find these criticisms compelling. I discuss them below and return to the alternative picture of rationality offered by Simon and Gigerenzer and colleagues in some detail in section 3.3. There have also been powerful internal criticisms of aspects of the picture, that is, criticisms from economists and from philosophers sympathetic to decision theory. I briefly review two of these criticisms first.

The core commitment of decision theory is that rational agents' preferences conform to the axioms. It is not surprising, then, that internal criticisms of decision theory focus on the tenability of some of these axioms. Amartya Sen has challenged the idea that a rational agent's preferences must be internally consistent (Sen, 1993). Another prominent criticism is that the theory as it stands does not take account of the fact that people's preferences are not all commensurable and thus cannot be put into any one preference relation.

Sen's attack on the limitations of the decision-theoretic picture of rationality is rather wide-ranging:

ence for *x* over *y*. Your personal utility is then defined as simply a numerical representation of this 'preference,' assigning a higher utility to a 'preferred' alternative. With this set of definitions you can hardly escape maximizing your own utility, except through inconsistency. (Sen, 1977, p. 322)

> A person is given one preference ordering, and as and when the need arises this is supposed to reflect his interests, represent his welfare, summarize his idea of what should be done, and describe his actual choices and behaviour. Can one preference ordering do all these things? A person thus described may be 'rational' in the limited sense of revealing no inconsistencies in his choice behavior, but if he has no use for these distinctions between different concepts he must be a bit of a fool. The *purely* economic man is indeed close to being a social moron. Economic theory has been much preoccupied with this rational fool decked in the glory of his *one* all-purpose preference ordering. To make room for the different concepts related to his behavior we need a more elaborate structure. (Sen, 1977, pp. 335–336)

One aspect of Sen's critique of consistency is the undeniable logical point that bits of behaviour are neither logically consistent nor logically inconsistent. Any set of propositions is consistent or inconsistent, but a set of choices does not by itself have any such property:

> The alleged requirements of 'internal consistency' are conditions that demand that particular internal correspondences hold between different parts of a choice function. The foundational difficulty with such conditions relates to the fact that choices are not, by themselves, statements that can or cannot be consistent with each other (Sen, 1993, p. 514).

There is more to the critique than this, however. Sen does not think that maximisation should be a consequence of the axioms of rationality. He is not opposed to the idea that people sometimes try to maximize their returns, but he argues that whether one seeks to maximize depends on one's aims, intentions and so on in a way that is sensitive to the context.

I do not want to go into all the details of Sen's critique, but one point worth making clear is that his work not only attacks the idea that having one preference relation which conforms to the axioms of decision theory is sufficient for rationality, as suggested in the quotation above, but also suggests that having such a preference relation is not necessary for rationality[106]. He argues against

[106] I suspect that Sen would not formulate his objections in quite this way, given his claim that

what could be called a methodologically behaviourist[107] or extensional economics, where all that is – or needs to be – known about a agent is the preference relation. Without knowing the *reasons* for an agent's preferences one does not know whether they should (rationally) prevail over those of other agents in case of conflict (Sen, 1976) or whether they should be internally consistent (Sen, 1993). Here I discuss Sen's criticism of the axiom of internal consistency of preference.

Internal consistency of choice may be formulated as 'Property α' (Sen, 1993), defined as follows:

(13) $x(\mathrm{S})$ and $x \in \mathrm{T} \subseteq \mathrm{S} \Rightarrow x(\mathrm{T})$, where S and T are sets of alternatives, and $x(\mathrm{S})$ means that alternative x is chosen from set S. (Gigerenzer, 2000, p. 202)

The formula in (13) says that if x is chosen from a set of alternatives, S, then it must also be chosen from a sub-set of S, T. This means that choice is insensitive to context and much else. As Gigerenzer comments, "No reference is made to anything external to choice – for instance, intentional states such as people's social objectives, values and motivations." (Gigerenzer, 2000, p. 203) Indeed this is the sense in which axioms of internal consistency are 'internal': "They are 'internal' to the choice function in the sense that they require correspondence between different parts of a choice function, without invoking anything outside choice (such as motivations, objectives, and substantive principles)." (Sen, 1993, p. 495)

"There is not much merit in spending a lot of effort in debating the "proper" definition of rationality." (Sen, 1977, p. 343)

[107] These is not Sen's term, but it is no exaggeration to use the term 'behaviourist', as his discussion of the history of the dominant interpretation of decision theory makes clear:

> Hicks ... became persuaded by the alleged superiority of the new [revealed preferences] approach, and warmly endorsed the study of human beings "only as entities having certain patterns of market behavior; it makes no claim, no pretense, to be able to see inside their heads".
>
> In the same spirit, Ian Little gave his stamp of methodological approval to this approach: "the new [Samuelson's revealed preference] formulation is scientifically more respectable [since] if an individual's behavior is consistent, then it must be possible to explain the behavior without reference to anything other than behavior". (Sen, 1993, p. 497)

Sen elsewhere explains the reasons for the rejection of reference to anything other than the choices an agent makes for understanding an agent's desires. An agent is assumed to reveal (to use the decision-theoretic term) his preferences by his choices:

> The rationale of this approach seems to be based on the idea that the only way of understanding a person's real preference is to examine his actual choices, and there is no choice-independent way of understanding someone's attitude towards alternatives. (Sen, 1977, p. 323)

The trouble with the axiom is that it seems to be false that agents have internally consistent preference orders. No consistent preference order can be given to someone who chooses *x*, rejecting *y*, on one occasion, but chooses *y*, rejecting *x*, on another occasion[108], unless we assume that the agent's preferences have changed between the two occasions. Of course, such an agent might simply have changed his mind or might not be a rational agent at all. The challenge for an opponent of property α is to show cases in which a rational agent whose preferences are stable, nonetheless fails to exhibit property α.

Sen (1993) gives examples where the addition of another alternative changes the situation so that it is intuitively plausible that an agent might choose differently. In one example, a diner offered *x*, an apple, or *y*, nothing, takes nothing because taking the last apple would be impolite. The addition of a second apple to the alternatives would have allowed the diner to take the original apple. This diner's choices contravene property α because he chooses *y* (nothing) over *x* (the apple) in the absence of *z* (a second apple), but *x* over *y* when *z* is an alternative. There is no good reason to think that this agent is irrational, but his preferences do not conform to the axioms of decision theory.

In an example given by Gigerenzer (2000), the presence of an alternative provides a clue to the agent about which situation he is in. The agent, again a guest at dinner, chooses nothing over snacks that are offered to him, anticipating a later offer of dinner. When he is also subsequently offered tea and cakes he chooses the original snack, since he infers that dinner will not be offered.

[108] By 'chooses *x*, rejecting *y*', I mean that the agent strictly prefers *x* to *y*. Obviously an agent who is indifferent between *x* and *y* could choose *x* on one occasion and *y* on another without contravening any axiom of consistency.

Again, the addition of an alternative – tea and cakes, this time – reverses the agent's previous choice. This agent, like the previous one, is rational, but does not come up to the standards required of rational agents by decision theory. Such examples provide strong considerations against the contention that conformity with the axioms of internal consistency of choice is a hallmark of a rational agent.

Examples of this kind are, effectively, thought experiments which demonstrate that axioms of internal consistency as formulated in decision theory are not necessarily applicable to the preferences of rational agents. As Sen shows (1993, p. 502), there are several kinds of factors which carry more weight than internal consistency. As well as "positional choice", illustrated by the apple example, and the "epistemic value of the menu", illustrated by the tea and cakes example, there are also cases in which rational agents exercise their "freedom to reject", as in fasting, exhibiting "a desire to violate, deliberately, the standard conditions of consistent behavior." (Sen, 1993, p. 502)

Another fundamental assumption of the decision-theoretic view of rationality is that an agent has a complete preference relation, where the desirability of any two alternatives is commensurable. This is the property of *continuity*. Contrary to the assumption, it is fairly clear that our preferences are not in fact all commensurable. This fact may be easier to accommodate within decision theory than it is to accommodate the previous objection that sets of preferences are not generally internally consistent. Nonetheless it is worth looking at this second criticism because it illustrates one of the central problems with classical theories of rationality: global consistency is not a plausible requirement.

The requirement of continuity can be expressed as one of the 'lottery' axioms: if there are three alternatives, x, y and z, x strictly preferred to y, which is strictly preferred to z, then there must be a lottery between x and z which is ranked equal as a choice with y. A lottery is just a list of outcomes (mutually exclusive, in this case), each with a probability. Thus if x is €1000, y is €500 and z is €0, an agent might accept that flipping a coin to decide between x and z is as desirable as simply receiving y, or he might prefer the chance of x to be higher, say 0.6 or 0.7. What is required by the axiom is that there is some probability $p < 1$, such that a lottery between x with probability p and z with probability $1-p$ is neither preferred to nor rejected in favour of y.

There are plenty of choices of x, y and z for which intuitively this is false. As Simon comments, "all of the available evidence seems to suggest that people do not have consistent utility functions, even at a single point of time, over all conceivable baskets of goods" (Simon, 2000, p. 37). Dreier (2004) gives as an example the choices in (14). In normal circumstances[109] no rational agent would accept a lottery in which instant death was one of the outcomes.

(14) x: gain a banana, no other change;

y: no gain or loss;

z: instant death.

Such examples demonstrate that rational agents do not typically have global preference relations over all outcomes. Indeed, with a bit of thought, it is easy to find pairs of outcomes which are both desirable but which belong to such different spheres that it is hard to know how to say which is preferable. I think that, for example, peace in Sudan and a postdoctoral position for me are both highly desirable, but I have no idea which I prefer. I suspect that I have no stable preference, and that attempts to get me to value them both in some common currency (money, for example) would fail because my preference would be context sensitive, depending on mood, the background information presented with the question and other considerations.

Dreier suggests that non-continuity could be accommodated by having different orders of goods. If this kind of solution were pursued, there would be a preference relation conforming to the axioms within each order. This amounts to partitioning the preferences, or to introducing extra dimensions of preference. This kind of partitioning concedes a great deal to bounded rationality, since it brings into the model a recognition that rationality does not require global consistency. Real rational agents do not make decisions by lining up all

[109] This is intended to exclude unusual circumstances in which the agent has such a compelling reason for preferring x to y that it is worth risking death to obtain x; also circumstances in which the agent rationally prefers death.

outcomes and going for the global maximum. We get by instead by making choices within limited areas. Since we rarely if ever have to choose between career advancement and world peace it does not matter if we have no settled or consistent preferences over such alternatives.

This discussion illustrates Simon's point that "the critical scarce factor in decision-making is not information but attention. What we attend to, by plan or by chance, is a major determinant of our decisions." (Simon, 1997, p. 124) We make choices from the limited range of options that we are considering at any moment. The reason for this is not a lack of information about the value of other choices, but the lack of abilities required to consider and weigh up simultaneously all the things that one could choose to do. The same goes for theoretical reasoning: we do not generate and weigh up all possible solutions to a problem. Does this mean that we miss good choices or good solutions? The answer is that it must, but not as much as one might imagine, since (a) the best solution, or at least a very good one, is often in the domain being considered; and (b) there are mechanisms which can make this kind of local search broader if necessary and if time allows. If none of the options is good enough, and there is time, none will be chosen and new candidates will be considered, perhaps from a different area or domain. If I cannot find a good flat in London after some effort, for example, I might give up on flats and consider other forms of shelter, or I might stop looking in London and start looking in Tokyo. I am unlikely to look in both places simultaneously.

3.1.2 CONSISTENCY OF BELIEFS

Philosophers have generally assumed that it is not rational to have logically inconsistent beliefs[110]. From any two inconsistent propositions, such as some proposition, p, and its negation, $\neg p$, any arbitrary proposition follows logically. In psychologically realistic terms, then, a danger posed by inconsistency is that a system for generating valid inferences, fed a contradiction as input,

[110] Without this assumption the much discussed 'preface paradox' (Makinson, 1965) loses its bite. The idea is that the common practice of acknowledging in the preface to a work that the work contains false statements, "appears to present a living and everyday example of a situation which philosophers have commonly dismissed as absurd; that it is sometimes rational to hold logically incompatible beliefs." (Makinson, 1965, p. 205)

may reach any conclusion whatever. Consistency is sometimes given as a minimum criterion for rationality, as in Elster's 'thin theory of rationality':

> Consistency, in fact, is what rationality in the thin sense is all about: consistency within the belief system; consistency within the system of desires; and consistency between beliefs and desires on the one hand and the action for which they are reasons on the other hand. (1983, p. 1)

In fact this criterion is very strong. As previously noted, there are two ways that a theory of rationality can be intended or taken. It can be either normative or descriptive. A normative theory says what a rational agent should do. A descriptive theory tells us what rational agents actually do, either in purely descriptive terms, or in terms of the natural laws that govern them as rational agents. There are strong reasons to think that real agents do have inconsistent beliefs. Anecdotally, it seems that we often hold beliefs that are inconsistent over considerable periods of time, perhaps discovering their inconsistency only when it is pointed out. It makes sense theoretically that we should have inconsistent beliefs, because we could not check the consistency of our belief systems even if we wanted to. As O'Brien says:

> ... ordinary reasoning would seem to have very little interest in assessing the consistency either of large premise sets or of large sets of potential theorems, and people often believe in contradictory propositions simultaneously without realising that they are doing so. (O'Brien, 2004, pp. 208-209)

The main theoretical argument that real agents do not check their belief systems for consistency is that there is no way that they could without computational explosion. Inconsistency is a property of a set of propositions as a whole. One can have direct inconsistency between two propositions in the set (e.g. p and $\neg p$). But there are also inconsistent sets of more than two propositions in which no two propositions are inconsistent[111]. Some of these sets are such that if any one proposition is removed from the set, the remaining propositions are consistent, as in examples (15) and (16):

[111] Cherniak calls such cases 'tacit inconsistencies' (1986, p. 16).

(15) A is taller than B.

B is taller than C.

C is taller than D.

D is taller than A. (Johnson-Laird, 2004, p. 191)

(16) If not A then B

If B then C

Not A and not C. (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000, p. 531)

Such examples establish that in order to ensure that a set of propositions such as a belief system is consistent, it is in principle necessary to check the whole set for consistency. The trouble with this is that checking a set of propositions for consistency is very computationally expensive for anything other than very small sets. Using the truth-table method to check the consistency of a set of *n* propositions, a table with 2^n rows is required (van Dalen, 2004, p. 20). An agent with only one hundred beliefs, checking ten complete rows per second, would take more than four thousand billion billion (4×10^{21}) years to check its belief set for consistency (see Appendix II). The fundamental problem, which cannot be finessed by more computationally efficient algorithms, is that the task of consistency checking grows exponentially with the number of propositions.

In practice, the need for global consistency testing is avoided by (a) segregation of beliefs, (b) the way that cognition is set up so that we are more likely to form and store true than false beliefs, and (c) the distinction between long-term and short-term memory. There is no general need to check that beliefs in different domains are consistent, since they are unlikely to interact, and our intuitions reflect that: "You know, or you think you know, that *this* belief has no bearing on *that* belief. Your belief, say, that George Bush won the 2000 presidential election is, you suppose, independent of your belief that water contains oxygen." (Johnson-Laird, 2004, p. 191) There is also no good reason to expend effort in checking the consistency of beliefs if one is reasonably confident that they are true, since all true propositions are, of course, consistent with each other. Given that consistency checking is prohibitively costly, one would expect evolved (or well-designed) rational agents to be set up so that they mostly avoid storing false beliefs in the first place. Human perception is mostly veracious, and propositions that come from inference, speculation or testimony from others can be subjected to limited consistency checking before being stored in long-term memory[112].

There are differences in what we expect from a rational agent as far as consistency is concerned, depending on whether the beliefs are in long-term or short-term memory, as Cherniak (1986) points out. It is a common assumption that there are two kinds of memory: long-term memory, a large-scale storage area, in which beliefs are stored but not acted on; and short-term memory, into which small amounts of information from the senses, from long-term memory and from inference is placed for short periods and in which active processing occurs. Given that long-term memory is inactive, some inconsistency is inevitable between beliefs in long-term memory. This is related to the fact that we do not strongly expect people to draw even obvious inferences from beliefs in long-term memory. If an agent knows $p \rightarrow q$ and later learns p, we are not certain that he will conclude q unless the belief $p \rightarrow q$ 'comes to mind', i.e. is retrieved into short-term memory.

Rational agents should draw obvious inferences from beliefs in short-term memory (Cherniak, 1986, p. 59), and therefore there is good reason for some consistency checking of these beliefs to avoid the drawing of arbitrary conclusions. Even for short-term memory, though, it is unlikely that consistency checking is exhaustive. Given a short-term memory that holds (e.g.) six propositions, the truth-table method requires a table with sixty-four rows. It is more likely that propositions in short-term memory are monitored for direct contradiction, so that if two propositions, one of which is the negation of the other, are present then the inconsistency is flagged and resolved. Sperber and Wilson's deductive device for spontaneous inference works like this (Sperber & Wilson, 1986, p. 95). Braine and O'Brien's mental logic has a rule that flags

[112] Testimony can also be assessed on the reliability of the source, but checking the internal consistency of what is asserted must often play a role. Sperber has even argued (2000; 2001) that the evolutionary function of reasoning ability is its use in evaluating others' assertions and the arguments they present to back them up.

such pairs as inconsistent and in addition a rule that registers inconsistency when confronted with pairs of propositions of the forms $p_1 \vee \ldots \vee p_n$ and $\neg p_1 \wedge \ldots \wedge \neg p_n$ (O'Brien, 2004, p. 212).

If the requirement to maintain a completely consistent set of beliefs is taken as normative, it is still questionable. Why should an agent eliminate all inconsistency? A standard answer might be: because it is irrational to believe something that one knows is false, and if one's belief set is inconsistent then the conjunction of all of one's beliefs is false. But it is not necessary to accept that if it is rational to hold each of the beliefs currently in one's belief set, then it is rational to believe the conjunction of one's beliefs (the 'Conjunction Principle'[113]).

A reason that might seem more pressing is the aim of eliminating the risk of deriving arbitrary conclusions, but there are other ways to avoid this danger. What is more, since it is impossible in practice for any being that works at a finite speed and has more than a handful of beliefs to check the consistency of its complete belief set, the norm would be unachievable. There are well-known philosophical problems concerning normative rules which it is completely infeasible to conform to. I do not know whether *'should' implies 'can'* is correct as a general rule, but it is at least worth bearing in mind. Perhaps the correct normative rule is something like: a rational agent should eliminate inconsistencies in his belief set when they might be harmful and can be detected without undue effort i.e. when it is likely to be worth doing.[114]

In the preceding sections I have discussed strong consistency requirements. I have indicated that there are strong links between global consistency and a requirement to optimize or maximize. In the next section, I discuss optimisation and maximisation in decision theory and in classical rationality in general.

[113] The Conjunction Principle has been discussed in the literature on the preface paradox, for example in Ryan, 1991; Douven & Uffink, 2003.

[114] Cherniak suggests a weaker 'minimal consistency condition' as a necessary condition for agenthood: "If A has a particular belief-desire set, then if any inconsistencies arose in the belief set, A would sometimes eliminate some of them." (Cherniak, 1986, p. 16)

3.1.3 OPTIMISATION, MAXIMIZATION AND CONSTRAINED MAXIMISATION

Maximizing expected utility is one way of optimizing. Roughly, maximizing means making some quantity as large as it can be: for example getting the highest return, or acting so as to get the highest expected return. Optimizing means doing the best that one can, so in a sense is broader than maximization.[115] Optimising might require simultaneously maximising several variables, for example. Sometimes the distinction is not clear or non-existent. The norm of classical rationality according to which all beliefs should be consistent can be seen as a requirement both to maximize and to optimize consistency.

Although philosophers have placed less stress on maximization of returns than decision theorists, many agree with economists that rational agents maximize returns, rather than satisfice, moved by considerations like the following: If one has reached a satisfactory outcome but could achieve more, it is rational to do so, all else equal. Sorenson pithily sums up this view: "... rationality demands opportunism. Imagine that you are well off but could double your fortune merely by lifting a finger. Is it rationally permissible to forego lifting a finger?" (Sorensen, 2004, p. 261) This question arises when one considers whether a rational agent can constrain his or her future behaviour in advance, if doing so would maximize overall returns, but only at the expense of adopting a principle that requires the agent to turn down the most desired option at some point in the future. (McClennen, 1990; Dreier, 2004, pp. 163–165; Sorensen, 2004, pp. 260–263)

One reason that this issue has been considered is that there are situations in which everyone does better (in the sense that all receive greater returns) if individual agents can resist the temptation to maximize at each moment. 'Prisoner's dilemma' situations are those in which both (or all) participants are better off when they both (all) cooperate with each other, than when they both (all) do not cooperate, but each agent is better off if he does not cooperate when the other one does. The name 'prisoner's dilemma' derives from the situation in which two criminals have been captured, both are facing impris-

[115] As I discuss below, Simon attacks optimization rather than maximization. I think that the reason is that optimization is the broader category.

onment, and both are offered a shorter sentence as an incentive to give evidence against the other. If only one gives evidence he is pardoned or receives only a token sentence, while the other gets the full sentence for the crime; but if both give evidence they both receive heavy sentences. If neither gives evidence then both will receive lighter sentences than if both confess (since it will only be possible to convict them of a lesser crime). According to standard game theory, a rational agent will not cooperate with the other prisoner in such a situation because non-cooperation (giving evidence) makes him better off if his opponent cooperates (keeps silent) and better off if his opponent does not. The apparent paradox is that both criminals would be better off if they both remained silent, receiving the short sentence for the lesser crime, but if they are rational by the standards of game theory and decision theory they will both confess because doing so maximizes expected utility.

A constrained maximizer (Gauthier, 1986) is an agent who acts according to a principle that, were it adopted by others, would make all better off. In games of prisoner's dilemma, a constrained maximizer will cooperate and may do much better overall than someone who maximizes at each moment (a straightforward maximizer). Some have argued that it is rational to be a constrained maximizer (Gauthier, 1986) since constrained maximizers do better overall in prisoner's dilemma situations. Others have rejected this claim on the basis that rational agents must decide what to do on the basis of expected, i.e. *future* return (Sorensen, 2004), or that a so-called constrained maximizer is really just an unconstrained maximizer who happens to prefer cooperating and is thus still behaving strictly in accordance with his aim of maximising his utility given his preferences when he does so (Dreier, 2004).

What advocates of constrained maximisation and of straightforward maximisation agree on, evidently, is the idea that rational agents maximize. The mainstream view is that the rational thing to do is to fulfil one's desires to the greatest extent possible, putting aside the issue of whether one's desires are rationally justified, that is. So the debate between constrained and straightforward maximizers is an internal dispute, in contrast to the more fundamental disagreement between bounded and classical rationality.

3.1.4 'AS IF' THEORIES OF COGNITION

Simon dubs optimising theories 'as if' theories of cognition, since they assume that agent's judgements and choices are as they would be if all solutions were somehow considered in the search and accurately assessed in the light of all relevant information so that the best is reliably found. As we have seen, advocates of optimization do not propose procedures that could perform this feat. Simon has a useful analogy: optimising solutions model problem solving as though it fitted the environment perfectly, like jelly being poured into a mould. If you want to know what shape the jelly will be once set, it suffices to know the shape of the mould. Similarly, to know what an optimizing system will do, it is not necessary to consider the properties of the system. An optimizing system will always find the best solution or solutions (the highest point once the jelly is turned out), so it is only necessary for the theorist to determine what these are in order to know what the system will do. This is how a great deal of work in economics has been carried out (with the notable recent exceptions of developing programmes of research in behavioural economics and in cognitive economics). As we have seen, decision theory is an 'as if' solution in that it treats agents as if they knew everything relevant about the problem of what to choose and took it all into account. The claim is that it is not necessary to consider how they do this, nor whether they do.

In sum, unbounded rationality suggests "building models that perform as well as possible with little or no regard for how time consuming or informationally greedy such models may be" (Gigerenzer, Czerlinski, & Martignon, 2002, p. 149). These models will therefore be poor models of human reasoning. The alternative is to "design models specifically to fit the peculiar properties and limits of the mind and the environment" (Gigerenzer, Czerlinski, & Martignon, 2002, p. 149), that is, to embark on the programme of bounded rationality.

If we assume bounded rationality, then we cannot assume that a system is powerful enough to consider the whole space of solutions. Therefore we have to consider how the space is explored: what is it that determines which solutions are considered, and in what order? The correct analogy then is not a jelly filling an indented surface, but a point object tracing a path across the surface (assuming search is serial: it would be two point objects for parallel search; several for multiply parallel search).

Then we have to answer the questions: what path is followed, and when does search stop? A stopping rule is needed because a search that has unlimited time is unrealistic – and amounts to optimisation, since the entire surface can be explored. What path it is best to follow depends on the structure of the environment. In the example given in the introduction of catching a ball, experienced catchers all attempt to follow a particular path (or rather one of a bundle of similar paths) through the problem space, and thus in this case also through real space.

A good deal of what counts in some problems may be the point at which the search is started: where on the surface the probe is positioned initially. As we will see, in some cases, the fastest and most frugal heuristics, the starting point is near enough to the stopping point, so search ends after one decision. In these cases, much of the work is done by recognition: recognition of the type of problem, and therefore which heuristic to apply, and recognition of the few important clues in the mass of available information. In other procedures, the path followed and the stopping rule are more important than where search starts.

I consider these questions about how to implement bounded rationality in section 3.3. Before this, I examine a considerable body of empirical evidence that human reasoning widely deviates from the norms of classical rationality.

3.2 EMPIRICAL EVIDENCE

3.2.1 OVERVIEW

As well as the strong theoretical considerations in favour of a view of rationality as bounded, there is considerable empirical evidence. In the psychology of reasoning and of judgement[116], a substantial body of research over the last four decades has established that participants in widely differing tasks give re-

[116] The usual division in the literature is between reasoning tasks – those which require logical deduction or abduction – and judgement tasks – which are intended to test abilities with probability and classification. There is also literature on 'choice', which might be called the psychology of economic decisions.

sponses that systematically deviate from logical and probabilistic norms of rationality. (e.g Wason, 1960; Wason, 1968b; Kahneman, Slovic, & Tversky, 1982; Evans, 1989; Manktelow & Over, 1993; Piattelli-Palmarini, 1994 is a popular survey; Shafir & Leboeuf, 2002 is a recent scholarly survey). (See section 3.2.2 below for descriptions of experimental tasks.) Participants give answers that seem to fly in the face of basic principles of logic and probability theory, reaching conclusions that do not follow from the information presented, and failing to take into account all of the evidence. The robustness of the results in the face of various debiasing techniques such as explicit instruction and reduction of cognitive load suggests that the explanation must lie at the level of competence, not performance.[117] This has been widely taken to have "bleak implications for human rationality". (Nisbett & Borgida, 1975, coined the phrase. Their work concerned base-rate neglect in probabilistic reasoning.) It is claimed that the results reveal pervasive mental biases best accounted for in terms of a strong tendency to use inappropriate non-logical rules or heuristics. (Tversky & Kahneman, 1974; Kahneman & Tversky, 1982; Kahneman & Tversky, 1996; Evans, 1972; Evans, 1984; Evans, 1989; Evans, 2006) There is a suggestion in the air, although not made explicitly by those in the heuristics and biases school, that the rules of logic and probability are not part of human reasoning competence.[118] This view, according to Gigerenzer, "has become the common wisdom in and beyond psychology" (Gigerenzer & Hoffrage, 1995, p. 684).

However, it has also been shown that performance on some reasoning tasks can be considerably improved by altering the format of the task without changing its logical form. Participants are apparently sensitive to the content and context of tasks. Some theorists have argued that this is because the format or subject matter of a task may call up dedicated mental machinery. A famous example is the proposal that there is a domain-specific adaptation for reasoning about social contracts (Cosmides & Tooby, 1992). Gigerenzer and colleagues (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Gigerenzer &

[117] Although Cohen (1981) argues that such conclusions logically cannot be drawn from such experiments.

[118] This opinion can be found in popular works, e.g.: "Tversky and Kahneman argue, correctly I think, that our minds are not built (for whatever reason) to work by the rules of probability" (Gould, 1991, p. 469).

Hoffrage, 1995; Hertwig & Gigerenzer, 1999) have proposed that working with probabilistic data in the form of frequencies evokes different concepts, mental models and calculations from those evoked by data encountered as percentages or fractions. Thus under specified circumstances, tasks which use frequency data will receive more accurate answers, possibly reflecting cognitive adaptation to the format in which probabilistic data were encountered during human evolution.

Gigerenzer and colleagues have also shown, building on work by Simon, that computationally simple heuristics can provide answers to some complex choice and judgement problems and that the strengths and weaknesses of the heuristics match well with human performance. These heuristics are fast and frugal, ignore much of the provided information, often involve canonically invalid shortcuts, do not obey classical constraints of consistency or transitivity, and satisfice rather than maximize.

A factual convergence between the view of rationality provided by the psychology of reasoning on the one hand, and by work on simple heuristics and on domain-specific abilities on the other, has been noted, for example by Samuels, Stich and Bishop (2002; see also Samuels & Stich, 2004). There is a consensus that to understand reasoning one has to look at the processes involved: the aim is "to understand the cognitive processes that produce both valid and invalid judgments" (Kahneman & Tversky, 1996, p. 582, cited with agreement in Gigerenzer, 1996, p. 592). These processes are often fast and simple and the results they produce cannot be predicted by assuming that cognition will find logically normative answers or a perfect match to the environment. Reasoning is, in a word, bounded.

A sign that the debate has largely been won by proponents of bounded rationality is that much discussion has shifted to other areas. One question which has attracted considerable attention over the last decade is whether there are two reasoning systems, one more classical and one 'quick-and-dirty'. It has been proposed that there is a correlation with the machinery used, so that analytical, normative reasoning is conscious and effortful, and distinct from non-canonical fast, subconscious (or unconscious) reasoning processes (Evans, 1984; Evans, 1996; Sloman, 1996; Stanovich & West, 1998; Evans, 2003). (See chapter 4.)

In a way, this latter work has been an attempt to see how bounded rationality arises, as well as, more obviously, to explain how it fits with our intuitions about logical norms. In various schools of thought it is now assumed that human reasoning is bounded and to be investigated in terms of the procedures and mental representations it employs. In the simple heuristics school, work continues on finding heuristics that apply to different tasks and finding common elements of heuristics (tools from an 'adaptive toolbox') that apply across domains (Gigerenzer & Todd, 1999; Gigerenzer, 2000). As discussed in the previous section, there are signs that economics is beginning to come to terms with bounded rationality (Conlisk, 1996 argues that it must). In contrast with these fields, cognitive psychology, since the inception of an information processing model, has always been an investigation into the properties of mental processes *as processes*, rather than into properties of their outcomes, as Simon (2000) points out. Work here is several decades deep[119], even if in the subfield of psychology of reasoning models of rationality are only now being adapted to fit. Psychologists have also, ironically, been more concerned than economists with the tradeoff between costs of decision making and accuracy (Conlisk, 1996, p. 671) (although plenty of psychological models are unbounded (Gigerenzer, 2004)[120]).

Disagreement remains over the appropriateness of answers that do not match normative criteria. Some commentators continue to view the results as a bleak indication that normative standards are not met. Thus in a review article, Shafir and Leboeuf claim that "research on reasoning has continued to document persistent and systematic shortcomings in reasoning abilities," (2002, p. 494) and "people often violate tenets of rationality in inadvisable ways," (2002, p. 491). There is agreement on the experimental results, but still considerable disagreement over their interpretation. Although Samuels et al.

[119] Although some processing models in which mental processes mirror classical norms have been proposed in recent decades, for example, mental logic for deductive reasoning (discussed in chapter 2) and broadly Bayesian learning mechanisms such as weighted associative networks as the basis of probabilistic judgements (e.g. López, Cobos, Caño, & Shanks, undated).

[120] "Optimization, with or without constraints, has also spread beyond economics. Psychologists often propose models of cognition that assume almost unlimited memory, storage capacities, and computational power. That is, many psychologists also build 'as if' models of behavior." (Gigerenzer, 2004, p. 391)

show convincingly that the heuristics and biases school, *pace* Gigerenzer, is not committed to the view that the rules of logic and probability are not part of human reasoning competence, real disagreements remain. On one side we have Shafir and Leboeuf: "People use intuitive strategies and simple heuristics that are reasonably effective some of the time but that also produce biases and lead to systematic error." (2002, p. 493). On the other is Gigerenzer:

> The study of cognitive errors has been dominated by a logical definition of errors. But this narrow norm tends to mistake forms of human intelligence that go beyond logic for stupid blunders, and consequently fails to unravel the laws of mind. (2005, p. 3)

The disagreement is not merely a matter of temperament and outlook. There is substantive disagreement over the appropriateness of applying context-independent norms to human reasoning. Gigerenzer has criticised research which is content to show systematic deviations from norms by eliciting responses which diverge from normative answers without proposing specific models of how participants reason. Without knowing what mental formats and processes are involved one cannot tell what rules (normative or otherwise) are being used. The mental processes and formats employed by participants depend on the format, context and content of the task, so research should be sensitive to these factors.

If the content and context of a task is recognised as important, then the participant's interpretation of the communicative acts involved must be seen as playing a fundamental role, as Cohen recognised (Cohen, 1981). The reason is that "content-blind norms overlook some of the intelligent ways in which humans deal with uncertainty, for instance, when drawing semantic and pragmatic inferences." (Hertwig & Gigerenzer, 1999, p. 275, see also Gigerenzer, 1996). The interpretation of the communicative acts involved in reasoning tasks is of fundamental importance to what the task is, and what conclusions are seen as worth deriving, and therefore to the performance of participants on the task (and to the assessment of their performance). Equipped with a pragmatic theory and some general assumptions about cognition – much of the work along these lines has used the tools of relevance theory – one can show, first, that reaching conclusions that are not deducible from the premises on their own is not illogical or irrational, and secondly, that it is to be expected that participants will infer conclusions that are relevant in preference to ones that are true but trivial or absurd, even if extra premises must be supplied to do so. These aspects of performance are, in particular, compatible with a theory that has it that reasoners seek relevant conclusions under the constraint that their conclusions are logically warranted by the presented premises together with some other information or principles. Thus we return to the Gricean themes that much reasoning involves unstated premises and that (some) reasoning, working fast, nonetheless aims at canonical validity.

3.2.2 BLEAK IMPLICATIONS FOR RATIONALITY?

As an example of the work in psychology of reasoning that has been seen as having bleak implications for rationality, consider first Wason's selection task (Wason, 1966; Wason, 1968b). Four cards are presented, for example those in figure 1, together with a conditional statement 'If a card has a 6 on the front it has an E on the back'.

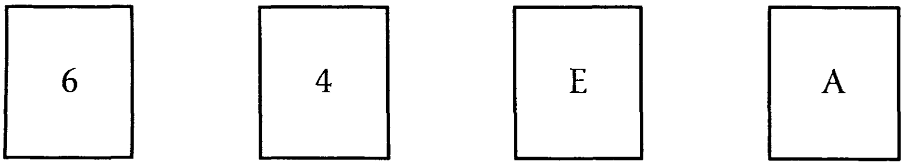

Figure 1: Wason selection task

The participant is asked which of the cards should be turned over to check the truth of the conditional statement. The normative response is *6* and *A*. If the card with a *6* on it does not have an *E* on the back then the proposed rule is falsified. The proposed rule has the structure: *If P then Q*. For the first card, we know that *P* is true (a *6* is printed on one side of the card), so if the rule holds, then by ***modus ponens***, *Q* must be true – there must be an *E* printed on the other side of the card. If there is no *E* on the other side then we have ***not-Q*** and we must give up *P* or *if P then Q* (or the rule of ***modus ponens***). We have the evidence of our eyes for *P*, so the proposed conditional must be given up. Similarly, for the fourth card – the card with an *A* on it – if there is a *6* on the other side then the rule cannot be true. To repeat, the rule is of the form: *If P* *then Q*. We know that the negation of *Q* is true, since there is an *A* on the card, not an *E*. Assuming the truth of the rule, then by *modus tollens* we have *not-P*. If the card actually has a *6* on it (*P*) then we have inconsistency and will have to drop our supposition that the proposed rule holds.

Another way of seeing the same point is to note that since the rule is of the form *If P then Q*, the only possible configuration that a card could have that is incompatible with the rule is *P and not-Q*. Thus the *P* card should be chosen, to see whether it has *not-Q* on the other side, and the *not-Q* card should be chosen to see whether it has *P* on the other side.

The normative answer is typically given by only a small percentage of participants. The rate at which it has occurred is not significantly different from the chance percentage of 6.25 (Noveck & O'Brien, 1996). The *P* card is chosen on the majority of trials, but the *not-Q* card is generally not chosen. More often the *Q* card is chosen, although it is logically irrelevant to testing the proposed rule: from *Q* and *if P then Q* neither *P* nor *not-P* follows, so whatever is on the other side of the card, this card will be compatible with the proposed rule.

These results have been reproduced many times with numerous variations on the task and materials. For the abstract version of the selection task (i.e. with abstract material such as letters and digits printed on the cards) the results have been rather robust[121]. Similarly poor performance has been observed in other experiments intended as tests of logical reasoning, such as the 2-4-6 task and relational problems (see below for descriptions). (For reviews see Manktelow, 1999; Johnson-Laird, 1999.)

Equally striking deviation from norms of rationality has been seen in experiments in which participants are asked to work with probabilities. In the Linda task, the so-called 'conjunction fallacy' is exhibited. Participants are given a description of a woman, Linda, and asked to rate various propositions concerning her in order of their probability. The relevant propositions are 'Linda is a bank-teller' (*A*), 'Linda is active in the feminist movement' (*B*), and 'Linda is a bank-teller and active in the feminist movement' (*A and B*). The de-

[121] The exception is the manipulations of relevance carried out by Sperber et al. (1995). The abstract form of the deontic-rule selection task often elicits normative answers, but this is a distinct task (see below), as Griggs and Cox (1993) argue.

scription does not entail any of these, but makes it clear that Linda's politics are liberal or progressive. Most participants rate the option *A and B* as more likely than *A* on its own. This ranking apparently violates the principle of probability theory that the probability of a conjunction of events cannot be greater than the probability of any one of the events: $P(A \text{ and } B) \leq P(A)$[122] (Kahneman & Tversky, 1973; Tversky & Kahneman, 1983).

The judgment-heuristics school of thought has stressed the negative implications of experiments like these for human rationality. (See Kahneman, Slovic, & Tversky, 1982, for the heuristics and biases programme.) The claim is that such results show that participants use simple non-logical heuristics, rather than normative rules of logic, to reach their judgements. For example, Tversky and Kahneman (1983) proposed that in the Linda task, the propositions are ranked according to a 'representativeness heuristic', where representativeness is a measure of the correspondence between the description and the proposition in question. The idea is that being a feminist is representative of the description given about Linda, whereas being a bank clerk is unrepresentative. Being both is somewhat representative and somewhat unrepresentative and is therefore ranked more likely than being a bank clerk but less likely than being a feminist.[123]

Generalising, the idea is that non-logical considerations such as the similarity between evidence and conclusions ('representativeness'), what comes to mind easily ('availability') and what is presented first ('anchoring'), collectively take precedence over such factors as the logical structure of the conclusions, their logical relation to the evidence and the confidence with which evidence is known. This kind of account has gained wide acceptance as an explanation for various deviations from norms in probabilistic reasoning, including participants' overconfidence (and occasional underconfidence) in their own judge-

[122] This law is a special case of a more general principle which can be seen as concerning implication or extension: if the extension of *X* is a subset of the extension of *Y*, i.e. *X* implies *Y*, then $P(X) \leq P(Y)$ (Politzer & Noveck, 1991, p. 90; Bonini, Tentori, & Osherson, 2004, p. 200).

[123] A similarly structured problem is as follows: estimate the probabilities of a) a flood somewhere in North America during 1983, in which 1000 people drown; b) an earthquake in California during 1983, causing a flood in which 1000 people drown. Since the *b* events form a subset of the *a* events, *b* cannot be more likely than *a*, but participants reliably rank *b* as more likely than *a*.

ments, overestimation of probabilities, and the neglect of base rates (for the last of these see below).

In a similar vein, Evans (1972; 1998; Evans & Lynch, 1973) gives an account of the abstract selection task in terms of a matching bias: a tendency to choose as answers the cards on which the symbols match the ones in the proposed rule. On Evans' account, the matching bias arises from the interaction of two heuristics: a 'matching-heuristic' that selects based purely on lexical similarity and an 'if-heuristic' that prefers material found in the antecedent of a conditional to material found in the consequent.

Evans' explanation has been attacked as little more than a redescription of the data[124]. Roberts characterises (but does not endorse) this view: "Why do people match? Because of the action of the matching heuristic. How do we know that a matching heuristic is applied? Because people show matching behaviour."[125] (Roberts, 2004, p. 248) In reply to such objections, Evans has maintained that the reasoning bias stems from, and should therefore be detectable in, an attentional bias to the matching cards: that "many [participants] decide first and think afterwards" (Evans, 1996, p. 238). He and Roberts and Newton have found some supporting evidence (Evans, 1996; Roberts & Newton, 2001), and Evans maintains that the matching-heuristic and the if-heuristic are predictive of behaviour and well-supported by experimental evidence on the selection task and other tasks (Evans, 1999).

Despite the proliferation of studies, there is no real agreement about the explanations for participants' choices on the selection task: "The selection task... has launched a thousand studies, but the literature has grown faster than knowledge" (Johnson-Laird, 1999, p. 127). There has been less consensus about the matching bias account of the selection task than about Kahneman and Tversky's judgment-heuristic approach in probabilistic reasoning. Alternative explanations have been given in terms of mental models (Johnson-Laird

[124] See also footnote 128 below for similar comments from Gigerenzer, aimed at work on heuristics in economics as well as in psychology.

[125] Compare with Nietzsche's objection to Kant:

> "How are synthetic judgements *a priori* possible?" Kant asked himself – and what really is his answer? By virtue of a faculty... But, is that – an answer? An explanation? Or is it not rather merely a repetition of the question? How does opium induce sleep? "By virtue of a faculty," namely the *virtus dormitiva*, replies the doctor in Molière ... But such replies belong in comedy... (Nietzsche, 1968, pp. 208–209)

& Byrne, 1991), and in terms of Bayesian calculations of the probability of falsifying instances, on the assumption that participants see the task in terms of data selection for inductive hypothesis testing (Oaksford & Chater, 1993).

Analyses in terms of mental models, heuristics and biases and Bayesian inference have been also been given for other logical reasoning tasks, again with no real consensus reached on the mental processes involved. However the mainstream opinion in psychology appears to be that the mental biases are real, and the pessimistic conclusion is often reached that the literature establishes that human reasoning disregards basic logic and goes wrong on simple deductive tasks[126]. Shafir and Leboeuf, for example, summarise the work on logical (as opposed to probabilistic) reasoning tasks thus: "All told, research on [logical] reasoning has continued to document persistent and systematic shortcomings in reasoning abilities" (2002, p. 494).

A more realistic assessment, in my opinion, is that the research helps to undermine the idea that human reasoning exhibits classical, unbounded rationality. Predictions of participants' responses based on the idea that they will simply conform to rules of logic or probability theory are generally wide of the mark. That is, one cannot know what answers people will give to reasoning problems without considering the way they reason. This is not to say that people reason poorly, or irrationally (or not at all). Rather, the content and the context of reasoning tasks can reasonably affect the way that participants tackle them.

As mentioned in the overview, one way in which these factors play a role is in their influence on the interpretation of the task. It is not just that participants may pragmatically infer logically richer premises on conversational grounds, but more generally that what participants do and the conclusions they reach may very reasonably depend on what they infer about the task that they have been asked to perform. I discuss this line of research, including Sperber and colleagues' convincing relevance-theoretic explanation of the selection task, below.

[126] This opinion is not necessarily shared by advocates of explanation of performance on logical tasks in terms of Bayesian reasoning. But they have not shown how such Bayesian inference is actually carrięd out: "there is, as yet, no corresponding theory of the mental processes underlying performance" (Johnson-Laird, 1999, p. 127)

A (compatible) observation is that to understand reasoning, and cognition generally, one needs to look both at the cognitive strategies employed by the mind and also at the structure of the environment. These are the two blades of Simon's scissors: "Human rational behavior (and the rational behavior of all physical symbol systems) is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor." (Simon, 1990, p. 7). Much work in the psychology of reasoning has focussed on the cognitive blade and neglected the importance of the match between the cognitive capacities used and the environment of the task. Thus the conclusion of bleak implications for human rationality has been drawn because the wrong criteria are used. Performance is compared with laws of logic or probability rather than the environment, but "to evaluate cognitive strategies as rational or irrational, one also needs to analyze the environment, because a strategy is rational or irrational only with respect to an environment, physical or social" (Gigerenzer, 2004, p. 397). Thus Simon's observation is of considerable importance for the psychology of reasoning, cutting the ground out from under the 'bleak implications' school of thought[127]. The observation also underpins Gigerenzer's simple heuristics programme, which I return to below. Here I consider a corollary, also of importance for the psychology of reasoning: "apparently stable cognitive illusions can be made to disappear and reappear by varying crucial structures of the environment." (Gigerenzer, 2004, p. 397)

Gigerenzer and colleagues and evolutionary psychologists have stressed that performance seen as problematic on reasoning tasks is often much improved when the information is presented differently, even if the task is formally equivalent. Distinct cognitive systems may have evolved for reasoning in different domains, that is, for dealing with input with certain types of content, in a particular context. Explanations along these lines have been offered for success on a deontic version of the selection task. Differences might also be due to adaptation to certain formats of information so that, for example, it may be easier to reason about frequencies of events than about probabilities, rather as long division is harder with roman numerals than with the familiar base-ten format (see Dehaene, 1999, p. 98ff, on the 'place-value principle').

[127] Gigerenzer calls it "the study of cognitive illusions and errors" (2004, p. 397).

It has been known for some time that frequency data can facilitate performance on probabilistic reasoning tasks. Teigen (1974) used formally equivalent questions asking the participant to estimate either the probability of a randomly chosen X (e.g. female student at the university of Bergen) being a Y (e.g. over 160 cm tall), or the number of Xs which are Y we would find if we checked a particular number (e.g. 500) of Xs, and found overestimation in the probability format but more realistic estimates with frequencies. Tversky and Kahneman (1983), early in their work on the conjunction fallacy, found that conjunction violations occur less on frequency judgements. Further research has shown that frequency formats can reduce incidence of the conjunction fallacy considerably (from around 80% to as low as 10%) (Fiedler, 1988; Hertwig & Gigerenzer, 1999). Gigerenzer, Hoffrage, and Kleinbölting (1991) found that overconfidence in one's own answers to general knowledge questions could be made to disappear completely when judgements of the likely number of correct answers were elicited rather than the probability that a particular answer is correct. Gigerenzer and Hoffrage (1995; see also Gigerenzer, 2000, ch. 7) found that participants' answers on Bayesian reasoning problems of the type that normally elicit neglect of base rates are closer to those regarded as normative when the probability data are presented as frequencies of events rather than fractions.

These results may reflect the facts that the natural way to learn about the probabilities of events is to observe a natural sample of events[128] and tally frequencies (Gigerenzer & Hoffrage, 1995), and that humans are relatively good at processing and storing frequency data and do so with little effort:

> A large literature suggests that (a) memory is often (but not always) excellent in storing frequency information and (b) the registering of event occurrences for frequency judgements is a fairly automatic cognitive process requiring very little attention or conscious effort (Gigerenzer, 2000, p. 137)

However this work goes further, proposing specific models and procedures for reasoning about confidence (Gigerenzer, Hoffrage, & Kleinbölting, 1991)

[128] A natural sample is one that is not chosen to include or exclude certain events, so that event frequencies should reflect underlying probabilities.

and for Bayesian reasoning (Gigerenzer & Hoffrage, 1995)[129]. Because of the explicit models, predictions can be nuanced and precise. Gigerenzer et al. (1991) predict under what conditions frequency judgements are more accurate than judgements of probability, but also when frequency judgements will be inaccurate, and explain why (Gigerenzer, 2000, p. 158). Gigerenzer and Hoffrage show how the calculations required for judgements of frequencies based on remembered numbers of events are computationally less demanding than those required for probability judgements. Given data in the frequency format, simple calculations produce Bayesian answers without the need to keep track of or mentally represent base rates (Gigerenzer & Hoffrage, 1995).

A number of theorists have proposed that humans are good at reasoning in specific domains. For example Cosmides and Tooby and colleagues (Cosmides, 1989; Cosmides & Tooby, 1992; Cosmides, Tooby, Fiddick, & Bryant, 2005) propose that we have an evolved domain-specific mental mechanism for detecting violations of social contracts – a so-called 'cheater detection' faculty. This has been taken to explain the much better performance seen with the selection task when it involves checking to see whether a social rule has been obeyed. Earlier work by Cheng and Holyoak and colleagues (Cheng & Holyoak, 1985; Cheng & Holyoak, 1989; Cheng, Holyoak, Nisbett, & Oliver, 1986; Kroger, Cheng, & Holyoak, 1993) offered a similar explanation in terms of domain-specific pragmatic reasoning schemas dedicated to reasoning about permission or obligation. Before Cheng and Holyoak's work, facilitation had been found with descriptive (as opposed to abstract) versions of the selection task, e.g. versions which used a realistic rule about events and put descriptions of events on the cards[130]. Cheng and Holyoak (1985) showed that

[129] Gigerenzer (1996) accuses the heuristics and biases school of failing to propose specific models of thought:

> the sheer proliferation of studies is not always identical to progress. An ever-larger collection of empirical results, especially results that seem to vary from study to study in apparently mysterious ways, can be more confusing than clarifying. If the psychology of judgment ultimately aims at an understanding of how people reason under a bewildering variety of circumstances, then descriptions, however meticulous and thorough, will not suffice. In place of plausible heuristics that explain everything and nothing – not even the conditions that trigger one heuristic rather than another – we will need models that make surprising (and falsifiable) predictions and that reveal the mental processes that explain both valid and invalid judgment. (1996, p. 595)

[130] The first study which found facilitation using descriptive content was Wason & Shapiro,

the facilitation was seen in an abstract deontic task but not in a descriptive non-deontic task. I give a typical descriptive, deontic task below. A further difference from Wason's task is that in deontic versions the task is typically to see whether a rule, known to be in force, is being observed, whereas Wason's task asks participants to discover whether or not a proposed rule is correct (Griggs & Cox, 1993; see also Girotto, Kemmelmeier, Sperber, & van der Henst, 2001).[131]

Most participants give the normative *P* and *not-Q* response (cards 1 and 4) to the following version of the selection task:

> imagine that you are a police officer on duty. It is your job to ensure that people conform to certain rules. The cards in front of you have information about four people sitting at a table. On one side of a card is a person's age and on the other side is what the person is drinking. Here is a rule: IF A PERSON IS DRINKING BEER THEN THAT PERSON MUST BE OVER 19 YEARS OF AGE. Select the card or cards that you definitely need to turn over to determine whether or not the people are violating the rule. (Griggs & Cox, 1982, p. 415)

Figure 2: Deontic selection task (Adapted from Griggs & Cox, 1982)

There may well be domain-specific reasoning capabilities of the types proposed by Cosmides and Tooby or Cheng and Holyoak. However the results from the deontic selection task and other reasoning tasks do not necessarily support this hypothesis because there is a relevance-theoretic explanation of greater generality. Relevance theory explains what factors make selections of

1971.

[131] Noveck and O'Brien (1996) carried out experiments in which the factors abstract/descriptive, reasoning from a rule/reasoning about a rule, and deontic/non-deontic were crossed.

each card more likely in both Wason's selection task and the deontic selection task (see section 3.2.4 below).

3.2.3 LIMITS OF THE EMERGING CONSENSUS

The heuristics and biases school, evolutionary psychologists and the simple heuristics programme all endorse the view that human reasoning is boundedly rational. I agree with Samuels et al (2002) that the pictures of human reasoning given by these different schools are largely congruent and that some apparent disagreements are mainly a matter of emphasis or rhetoric. The heuristics and biases school's claim that "people's intuitive judgements on a large number of problems ... regularly deviate from appropriate norms of rationality" (Samuels, Stich, & Bishop, 2002, p. 240) is indeed entirely compatible with the claims of the opposing school that "there are many reasoning problems ... on which people's intuitive judgements *do not* differ from appropriate norms of rationality." (Samuels, Stich, & Bishop, 2002, p. 244). There is also some consensus that the format or content of problems can affect the accuracy of the treatment that they receive.

As Samuels et al. note, there remain serious disagreements about human rationality which hinge on the correct interpretation of probability theory, which is outside the scope of this thesis[132]. However I have tried to show that the consensus breaks down at another important point. That is the question of whether there is only one correct (i.e. rational) answer to the problems studied. Psychologists of reasoning generally assume that there is a unique norm-

[132] Samuels et al. discuss the role that differing theories of probability play in disagreement between Gigerenzer and Kahneman and Tversky. Kahneman and Tversky treat single-event probabilities as respectable theoretical entities. Gigerenzer endorses the frequentist interpretation of probability theory on which single-event probabilities are nonsensical. Samuels et al. say that (regardless of which is the correct interpretation of probability theory): "evolutionary psychologists cannot comfortably maintain both (a) that we don't violate appropriate norms of rationality when reasoning about the probabilities of single events and (b) that reasoning improves when single event problems are converted into a frequentist format." But I think that this is wrong. A frequentist could continue to assert (a) on the grounds that there are no appropriate norms which validly apply to single events, while claiming (b) that a frequency format improves reasoning, on the grounds that reasoning about single event probabilities is simply a confused attempt to reason about real probabilities (in the frequentist sense) and that translating the problem into a frequentist format facilitates such reasoning.

ative answer to each reasoning problem. Gigerenzer is more concerned with the rationality of cognitive strategies than answers. Whether an answer is rational depends on whether the strategy or heuristic that produced it is rational, and rationally applied. According to this school of thought, good performance on reasoning problems is a matter of having the right tool for each kind of task in one's toolbox and using it appropriately. Some experimental tasks do not test the abilities they have been thought to probe. There is a sense in which the correct answer to Wason's selection task is to pick the *P* and *not-Q* cards. But Sperber et al.'s work on the selection task shows that there are other cards which participants can pick, working on perfectly rational assumptions.

Before moving on to look at this and related work on the interpretation of tasks, I want to note that at least one so-called bias is best seen as evidence of participants' possession of rationality and common sense. When participants are presented with a complete syllogism and asked to evaluate the proposed conclusion on the basis of the premises presented, participants are more likely to endorse conclusions that are believable on the basis of general knowledge (e.g. "some babies cry") than conclusions that are unbelievable on that basis (e.g. "no babies drink milk"). It might seem that this is a failure of rationality: after all, as the task is set up, only the logical relationship between the premises and the proposed conclusion is relevant. Things are not so simple. It is rational to make use of general knowledge when evaluating conclusions in certain circumstances. One such circumstance is when a proposed conclusion does not follow with necessity from the premises provided, but is compatible with them. Here, even if the premises are accepted, the premises provide insufficient information to decide whether the proposed conclusion is true, and it is then perfectly rational to turn to general knowledge in deciding whether to endorse it. On the other hand, if the proposed conclusion is incompatible with the premises, then it can be ruled out on that basis. It turns out that the belief effect is much larger when the conclusion is possible but not necessary relative to the premises than when it is incompatible with the premises. Indeed when the conclusion to be evaluated is logically incompatible with the premises the belief effect has been found to be very small or not present at all (Newstead, Pollard, Evans, & Allen, 1992).

An intermediate case is when the proposed conclusion follows with necessity from the premises. It could be accepted on that basis, but it is plausible that at least sometimes people would apply a believability filter, rejecting a conclusion on the basis that its clash with general knowledge is more important than its following from the premises supplied (which may themselves be implausible). So if participants decide rationally then there should be a large belief effect when the premises neither necessitate nor rule out the conclusion, and a smaller belief effect when the premises necessitate the conclusion. That is exactly what has been found to be the case (Evans, Barston, & Pollard, 1983). One could only say that these results support the claim that there is a belief *bias* if one assumes that the best description is that participants are doing poorly on the (rather obscure) task of judging what follows with logical necessity from arbitrary premises. The more obvious description is that it shows that there is a perfectly rational belief *effect*. Participants take the task to be the commonly encountered one of assenting to a conclusion (or withholding consent). They work out the logical relationship between the premises and the conclusion, and bring general knowledge to bear in precisely the situations when it makes sense to do so.

3.2.4 PRAGMATICS AND THE PSYCHOLOGY OF REASONING

> Semantic inferences – how one infers the meaning of polysemous terms such as *probable* from the content of a sentence (or the broader context of communication) in practically no time – are extraordinarily intelligent processes. They are not reasoning fallacies. No computer program, to say nothing of the conjunction rule, has yet mastered this form of intelligence. Significant cognitive processes such as these will be overlooked and even misclassified as "cognitive illusions" by content-blind norms. (Gigerenzer, 1996, p. 593)

There is now a considerable body of work showing that pragmatic factors play a role in reasoning experiments, including work by Politzer (1990; 2005) on reasoning from statements containing quantifiers; by Schwarz, Strack, Hilton and Naderer (1991), Macchi (1995) and Politzer and Macchi (2005) on reasoning with base rates; by Dulany and Hilton (1991) and Politzer and Noveck

(1991) on the Linda problem; by Politzer & Nguyen-Xuan (1992) on reasoning about conditional promises and warnings; by Sperber and colleagues (Sperber, Cara, & Girotto, 1995; Girotto, Kemmelmeier, Sperber, & van der Henst, 2001; Sperber & Girotto, 2002) on the Wason and deontic selection tasks; by van der Henst, Politzer and Sperber (2002) on relational problems; and by van der Henst (2006) on the '2-4-6' problem. (Hilton, 1995; Politzer, 2004 are general papers). The need to look into how reasoners interpret the tasks they are given was identified by Cohen (1981). As Cohen says:

> it is always necessary to consider whether the dominant responses given by subjects in such [reasoning] experiments should be taken, on the assumption that they are correct, as indicating how the task is generally understood – instead of as indicating, on the assumption that the task is understood exactly in the way intended, what errors are being made. (Cohen, 1992, p. 419)

Psychologists have long been aware that one can only show that mistakes in reasoning are being made if participants' mental representation of the information given is as intended. The contribution that utterance interpretation makes to the way a participant represents a task has not always been appreciated, however (Hilton, 1995).

At the most straightforward, the interpretation of the information presented may involve pragmatic enrichment, so that participants are not necessarily reasoning from some kind of literal, bare-bones interpretation of the information explicitly given in the task. Politzer and Noveck (1991), show that on Gricean or relevance-theoretic grounds, participants in the Linda task might enrich the *A* response ('Linda is a bank teller') to *A and not-B* (Linda is a bank teller and not a feminist). Thus when these participants rate the *A* state as more probable than the *A and B* state, they are actually saying that *A and not-B* is more probable than *A* and *B*. This does not contravene any rules of probability.

This sort of consideration is now acknowledged in the literature, so that researchers wanting to demonstrate mental bias now attempt to factor out variant interpretations, for example by explicitly including an *A and not-B* choice in conjunction problems on the assumption that that would make it "pragmatically impossible" to interpret the *A* choice as *A and not-B* (Tentori, Bonini, & Osherson, 2004). While the manipulations attempted may sometimes be pragmatically naive[133], it is clear that variation of interpretation depending on the circumstances of communication is at least recognised as a factor.

What has not been so widely appreciated thus far is that subtler factors to do with communication also need to be considered. On Gricean grounds, communicative acts carry a presumption that they will meet certain standards. Grice set out such standards in the conversational maxims. Utterances should be truthful, informative, perspicuous, relevant and so on. In relevance theory, similar work is done by the presumption of optimal relevance. Many reasoning tasks are pragmatically odd from this point of view. (See the discussion of van der Henst's work on the '2-4-6' problem, below.) For example, in conjunction problems, to find the normative answer the participant needs to realise that only the form of the answer matters, and *A* should be rated more probable than *A and B*. But it is pragmatically odd that an interlocutor would go to such lengths as in the Linda task to convey a description that she knows is of no relevance to the matter in hand, so the task is systematically misleading. Studies that ask participants to bet repeatedly on options with different content but always of the form *A* and *A and B* (Sides, Osherson, Bonini, & Viale, 2002) – or *A* and *A and B* and *A and not-B* (Tentori, Bonini, & Osherson, 2004) – may be even more pragmatically strange. The normative response is to ignore all of the content and bet on the *A* option all down the line. But the act of uttering all of the content of the questions raises the pre-

[133] For example, the paper cited in the text does not explain what makes an *A and not-B* interpretation pragmatically impossible, and offers an odd choice between three options that are not mutually exclusive. Compare with the distinctly strange, *??Do you own a) a bicycle; b) a bicycle and a car; c) a bicycle and no car?*

Sides, Osherson, Bonini and Viale (2002) assumed that the *A and not-B* interpretation could be suppressed by telling participants that their chosen response would be shown to an independent judge who could not read the other response. Since the judge sees only *A* when *A* has been chosen the judge will not have pragmatic grounds for an *A and not-B* interpretation. Sides et al. optimistically assume that subjects will work this out and that this will influence their own interpretation to the extent that they do not interpret *A and not-B* as *A*.

Both of these studies have a further pragmatic oddity. They ask the participants to place bets on a series of choices, where the normative answer is to ignore all of the content of the choices and bet repeatedly on the *A* option. See the main text for discussion of this point.

sumption that it must be relevant to the task, rendering it unlikely that participants will simply ignore it. Unsurprisingly, both studies found that almost no participants bet only on *A* options.

Even some work that has focussed on the effects of utterance interpretation has understated its potential influence. *Pace* Hilton, the influence of utterance interpretation systems on performance in reasoning tasks need not be confined to a "front end component that determines how the incoming message is interpreted in its context" (Hilton, 1995, p. 249), or at least in some cases this 'front end' may do all of the work. On some tasks, the mental systems devoted to utterance interpretation may totally pre-empt domain-general or domain-specific reasoning systems. Sperber and colleagues have demonstrated that performance on the selection tasks (deontic and non-deontic) can best be understood – and manipulated – in this way.

Sperber, Cara and Girotto (1995) argue that what underlies successful performance on the selection task is not a domain-specific faculty such as a cheater detection mechanism or a pragmatic permission schema but pragmatic factors affecting interpretation of the conditional statement. They state that what matters is the way that the proposed rule achieves relevance.

Relevance theory is a general theory of cognition which defines relevance as a property of inputs to cognitive processes. Recall that the relevance of an input is a positive function of the cognitive effects achieved by processing it and a negative function of the effort required to process it. In the case of ostensive-inferential communication, utterances create a presumption of optimal relevance: the hearer is entitled to assume that an utterance is at least relevant enough to be worth processing, and what is more, is the most relevant one compatible with the speaker's abilities and preferences. This means that the hearer is justified in following a path of least effort in deriving the explicit meaning and implications of an utterance, stopping when an interpretation has been reached that satisfies his expectations of relevance. This is the relevance theoretic comprehension procedure. (This approach to ostensive inferential communication is set out in Sperber & Wilson, 1986; Sperber & Wilson, 1995; the term 'relevance-theoretic comprehension procedure' was introduced in Sperber, Cara, & Girotto, 1995, and discussed in Sperber & Wilson, 1995.)

From the definition of relevance as a positive function of cognitive effects and a negative function of processing effort, it follows that in an experimental situation, different interpretations can be made more likely by manipulating the effects which will be achieved by deriving a particular conclusion or the effort a participant will need to expend to derive it.

Returning to the selection task, a conditional statement of the form *if P then Q* has a number of derivable consequences including the following: that the consequent *Q* will be true when the antecedent *P* is true; that *P* and *Q* will be true together; and that *P* and *not-Q* will not be true together. Choosing cards on the basis of these interpretations leads respectively to selection of the *P* card only (6); the *P* and *Q* cards (6 and E); or the *P* and *not-Q* cards (6 and A) (see figure 1). To make it likely that participants make the normative choice of the *P* and *not-Q* cards the corresponding interpretation must be more relevant then the others in the context. In most contexts this is not the case, but by manipulating the effort and effects involved Sperber et al. were able to obtain a majority of correct responses. The successful scenario involved a card-printing machine which is supposed to comply with the conditional statement 'If a card has a *6* on the front it has an *E* on the back' but which had malfunctioned, printing *A*s instead of *E*s. Here the conditional statement becomes relevant by implying that the machine will no longer print cards with a *6* on one side and an *A* on the other. In this scenario, as predicted, these cards were preferred.

In a further experiment, Girotto, Kemmelmeier, Sperber and van der Henst (2001) showed that participants could be induced to select the *P* and *Q* cards on the deontic selection task by varying the scenario to make it relevant to find instances of compliance with the rule rather than rule violation. They also reproduced Sperber et al.'s results with the non-deontic task with new content. These results demonstrate that the kind of reasoning that is decisive in the selection task uses neither domain-general reasoning abilities, nor domain-specific abilities of the kind proposed by Cosmides and Tooby or Cheng and Holyoak. Instead, the mental apparatus which deals with ostensive stimuli appears to be used.

Even on types of task that do bring non-pragmatic reasoning systems into play, pragmatic factors will have a strong influence on the expectations and goals that participants have for a task, and thus on whether they will follow their reasoning through to a conclusion or consider any conclusion reached worth reporting. For example, van der Henst, Politzer and Sperber (2002) have shown that relevance theory successfully predicts when participants in indeterminate relational problems will respond that nothing follows from the information presented.

As mentioned above, many of the tasks in the reasoning literature, looked at from the point of view of pragmatics, turn out to be seriously misleading. Van der Henst (2006) argues that this is the case in the 2-4-6 problem and that participants' behaviour is best explained in terms of their interpretation of what is communicated by the experimenter. The task (Wason, 1960) is supposed to elicit reasoning that proposes and tests hypotheses. The experimenter asks the participant to find out what rule is obeyed by sequences of three numbers. The experimenter starts off the investigation by saying that the sequence '2, 4, 6' obeys the rule. The participant is invited to propose further triples to test the rule. Many participants infer rules such as *consecutive even numbers*, or *arithmetical progressions that increase by 2*, or *sequences of even numbers, increasing in size*, and only propose sequences that obey these rules. Thus they fail to discover that the rule is simply *numbers (monotonically) increasing in size.* Several explanations for participants' responses have been proposed. One, the so-called 'confirmation bias', is that participants fail to see that they should attempt to falsify the hypothesis they have in mind (Wason, 1960; Wason, 1968a). A second is that that they attempt to falsify but choose suboptimal triples – 'positivity bias' (Evans, 1989). A third is that they can only consider an unsuitably limited range of hypotheses – 'restrictiveness bias' (Poletiek, 2001). Despite disagreements about the mechanisms involved, the literature has generally considered participants' responses to demonstrate inadequacy in reasoning.

Van der Henst convincingly argues that the way the task is set up, providing the sequence '2, 4, 6' is misleading.[134] There are some highly salient properties of this triple: the numbers are the three smallest even numbers in order of size; and they are the first three numbers in the two-times table. Furthermore, the triple is part of an utterance which is, as an utterance, presumed by

[134] The task has been criticised as misleading since Wetherick, 1962.

the participant to be optimally relevant; the experimenter is assumed to be knowledgeable and trustworthy; and the task is to discover a rule. Thus, "any rule-like property that easily comes to mind when processing the initial triple should be considered by the participant as one the experimenter wanted him to consider in order to discover the rule." (van der Henst, 2006, p. 236)

Information provided by an interlocutor is different in this regard from information gleaned from the environment. Such information, because it is not communicated, does not come with a presumption of optimal relevance. Scientists are suspicious of overly neat data, considering the patterns likely to be coincidental and attempting to falsify the most obvious hypotheses. It is no surprise that van der Henst and collaborators found that when participants saw the triple generated by what they were told was a 'random' number generator, they acted more like scientists. They found the correct rule more quickly and more often than in the standard, communicative condition, proposing more triples that were not arithmetical progressions or did not increase.

It seems that at least some, and perhaps a great deal, of what has been taken to be accomplished by mechanisms dedicated to human reasoning, either domain-general or domain-specific, relies on the mental machinery for understanding utterances. When non-normative answers are given in reasoning experiments, pragmatic factors must be considered before the conclusion is reached that mental biases are in evidence, since the task that participants are attempting is likely to be different from what the experimenters think it is. This does not mean that cognitive biases do not exist, but in the absence of pragmatic analysis, these biases can easily be overestimated.

In section 3.1 I discussed theoretical considerations which suggest that human reasoning ability must be bounded – limited by the finiteness of human processing power – so that exhaustive consistency checking and exhaustive search are both impossible. This undermines classical models of rationality, which hold up optimisation and global consistency as norms and as approximate descriptions of human capabilities. In the current section, I have reviewed experimental evidence that has been used to argue that we entirely lack reasoning competence, finding that such a drastic conclusion is not justified, although there is plenty of evidence that human reasoning is neither unbounded nor insensitive to context. A more plausible account of the evidence involves taking the view that we are capable of making inferences that are valid, but since it is hard for us to discount information that seems relevant, including conversational clues about the task in hand, the conclusions we reach often differ from those that would be reached by a purely analytic approach.

In the next chapter I return to the pragmatic faculty as an object of study in its own right. Before that, in the final section of this chapter, I look at the positive programme of bounded rationality, with its stress on understanding the procedures involved in rational activity and the way that they exploit features of the environment.

3.3 BOUNDED RATIONALITY AND HEURISTICS

> Bounded rationality is simply the idea that the choices people make are determined not only by some consistent overall goal and the properties of the external world, but also by the knowledge that decision makers do and don't have of the world, their ability or inability to evoke that knowledge when it is relevant, to work out the consequences of their actions, to conjure up possible courses of action, to cope with uncertainty (including uncertainty deriving from the possible responses of other actors), and to adjudicate among their many competing wants. Rationality is bounded because these abilities are severely limited. Consequently, rational behavior in the real world is as much determined by the "inner environment" of people's minds, both their memory contents and their processes, as by the "outer environment" of the world on which they act, and which acts on them. (Simon, 2000, p. 25)

3.3.1 INTRODUCTION

The study of bounded rationality starts from Simon's claim that:

> It is impossible for the behaviour of a single, isolated individual to reach any high degree of rationality. The number of alternatives he must explore is so great, the information he would need to evaluate them so vast that even an approximation to objective rationality is hard to conceive. (Simon, 1997, p. 92).

In section 3.1, I looked at the theoretical debate between the bounded and unbounded visions of rationality. The essential points are summarised in Simon's list of ways in which what he then called 'objective' rationality (i.e. classical, unbounded rationality) is an unrealistic idealisation:

> Actual behaviour falls short, in at least three ways, of objective rationality... :
>
> (1) Rationality requires a complete knowledge and anticipation of the consequences that will follow on each choice. In fact, knowledge of consequences is always fragmentary.
>
> (2) Since these consequences lie in the future, imagination must supply the lack of experienced feeling in attaching value to them. But values can be only imperfectly anticipated.
>
> (3) Rationality requires a choice among all possible alternative behaviours. In actual behaviour, only a very few of all these possible alternatives ever come to mind." (Simon, 1997, p. 93)

Simon's comments are focussed on the rationality involved in choice of courses of action. In the case of theoretical rationality, similar considerations apply. Unbounded rationality would require that in making a judgement, one consider at least the following[135]:

(1) All possible solutions to each problem.

(2) All information that might be relevant in that it supports or undermines any possible solution. For non-demonstrative inference, absolutely any information[136] might be relevant. (These first two points amount to what Fodor calls *isotropy*.)

[135] One might also need to consider the best evaluation procedure for a judgement, and whether to start investigation at all. Each of these considerations might lead to an infinite regress: how to decide how to decide etc. what is the best evaluation procedure or whether to start investigation. See discussion in the main text below.

[136] There can be no stopping point for information search, since consulting all information known to the thinker is not enough – the environment is also full of information, all of which might have to be brought to bear on any judgment. To find truly optimal solutions a thinker in principle has to consider information from all sources, including libraries, advertisements, the internet and the memories of other people. (Sperber & Wilson, 1996, p. 530. See also Todd & Gigerenzer, 2000, pp. 729–730. On the role of adverts, see Stigler, 1961.)

(3) The consequences for one's belief system of the candidate belief, including such global properties as overall simplicity and consistency. (This is the claim that central thought is *Quinean*, in Fodor's terms.)

In a representational-computational model of cognition, it is impossible that these factors could be taken into account for each inference or judgment. Each one of them on its own would give rise to a computational explosion. As discussed in chapter 2, for this reason, Fodor, committed to the Representational Theory of Mind, concludes that we have no idea how central cognition works since there is apparently no way to model abductive reasoning in terms of classical computations (Fodor, 2000, p. 77). In other words, the criteria that Fodor insists on for central cognition and abductive reasoning amount to a view of rationality as unbounded, leading to pessimism about modelling it computationally.

We have seen that another way to proceed if one shares an unbounded view of rationality is to downplay the importance of understanding how human cognition works, and assume that it finds optimal solutions, *as if* it could take all of these factors into account, "building models that perform as well as possible with little or no regard for how time consuming or informationally greedy such models may be" (Gigerenzer, Czerlinski, & Martignon, 2002, p. 149). I have provided some theoretical arguments against this approach and discussed some of the considerable empirical evidence that human reasoning ability is not unbounded in sections 3.1 and 3.2.

The argument is incomplete without the presentation of an alternative. In this section I attempt to fill in the essentials of a programme modelling bounded rationality through search guided by heuristics. There are several key ideas to this approach: the inevitability that rationality is bounded; the idea of problem solving through generation and evaluation of trial solutions; the use of heuristics to guide search; the role of stopping rules, satisficing and aspiration levels; and the fit between the environment and the mind. Before considering these issues I look at an alternative, which is sometimes regarded as a (or the) form of bounded rationality, but is as psychologically unrealistic as unbounded rationality: optimisation under constraints.

Optimisation under constraints, originating in the work of Stigler (1961), keeps the ideal of optimisation, but factors in as constraints the costs of search for information. This is a step towards realism in that it does not assume that all agents are perfectly informed about what choices exist, their benefits and so on. The aim is to model solutions reached without consultation of all possibly relevant information, in contrast to unbounded rationality. This approach is referred to by practitioners as 'bounded rationality' (e.g. Sargent, 1993) (to Simon's great annoyance (Gigerenzer, 2004, p. 391)).

Because this work keeps the requirement to optimize, it is no more computationally realistic than previous models: indeed, "optimization under constraints can require even more knowledge and computation than unbounded rationality" (Todd & Gigerenzer, 2000, p. 730, for this point see also Winter, 1975; Vriend, 1996[137]) – because now an optimal stopping point must be calculated, and this is computationally intensive, the more so the more constraints there are. It is simpler to assume that everyone knows everything than to model ignorance and its effects. As Sargent says, "Ironically, when we economists make the people in our models more 'bounded' in their rationality ... we must be smarter, because our models become larger and more demanding mathematically and econometrically." (1993, p. 2) But in psychologically realistic explanation these costs must fall on agents, requiring "that the mind should calculate the benefits and costs of searching for each further piece of information and stop search as soon as the costs outweigh the benefits." (Todd & Gigerenzer, 2000, p. 729)

Calculation at each stage of the costs and benefits of continuing the search would again lead to a computational explosion: "the paradoxical approach is to model 'limited' search by assuming that the mind has essentially unlimited time and knowledge with which to evaluate the costs and benefits of future information search." (Todd & Gigerenzer, 2000, p. 730) The problem faced by this kind of agent includes and is worse than the original problem: "Taking account of the fact that decision-making is a costly activity necessarily leads to a more complex, recta-optimization procedure that includes the basic decision problem plus the problem how many costly resources to allocate to that original problem." (Vriend, 1996, p. 278).

[137] Vriend takes arguments against optimisation under constraints to be arguments against the programme of bounded rationality, wrongly thinking that it falls foul of the regress discussed in the text.

As Vriend recognises, this is a recurrence of a more general issue: how do we decide what to investigate? If by investigation, an infinite regress looms. One well-known expression of the problem is due to Ryle: "Must we ... say that for [an agent's] reflections how to act to be intelligent he must first reflect how best to reflect how to act?" (Ryle, 1949, p. 31).

To avoid the infinite regress, the general answer to Ryle's question has to be negative.[138] However, this argument provides no reason to think that cognition cannot involve a small number of layers, the earlier ones feeding the later ones with problems and information.[139] In models of bounded rationality, simple categorising and recognition mechanisms narrow down the search space in which simple reasoning procedures operate, as I discuss below. Here the regress is not infinite and there need be no computational explosion as long as the procedures are individually frugal.

3.3.2 THE INEVITABILITY OF BOUNDED RATIONALITY

Work on bounded rationality has mostly been concerned with heuristic processes (also known as heuristics): useful but not infallible shortcuts. Before examining the concept *heuristic* and the ways that heuristics are used, it is worth noting that heuristics are not a logically essential component of a theory of bounded rationality. Assuming limits on processing abilities, such as capacity limitations on working memory and the finite speed of information retrieval and processing, even a system that used only infallible (algorithmic) procedures would exhibit bounded rationality, since the algorithms used might take more resources than are available, and because already limited time, effort and processing capacity must also be divided between tasks.

Consider, for example, a long division task, such as dividing 10,934 by 345. I know an algorithm for tasks of this type, but it takes time and concentration,

[138] In my opinion Ryle was in not in a good position to support his own negative answer to his question, since he opposed psychologically realistic views of cognition. A commitment to cognition as computation recasts this regress as a form of computational explosion.

[139] Cherniak proposes a solution of this type to Ryle's regress, postulating:

> non-conscious mechanisms of selection or guidance that do not involve reasoning processes of any kind. These mechanisms may be acquired – for instance, as learned 'cognitive styles' – or the agent may be 'designed' by natural selection so that, as an efficient organism, he undertakes particular inferences. (Cherniak, 1981, p. 169)

so if these are not available or if other tasks require my attention, then I will not find the correct answer that way. An agent with limited processing resources and a long-division algorithm might not produce an answer at all or might produce the wrong answer because of a performance error (e.g. memory overload causing failure to 'carry' a digit).

Heuristics have been the subject of investigation because of their potential to make the best of limited resources. For example, I might reason heuristically: 345 is a bit bigger than 333 and 10,934 is a bit smaller than eleven thousand, so the answer will be slightly less than 11,000 divided by 333. Three-hundred-and-thirty-three goes into one thousand about three times; and there are (obviously) eleven thousands in eleven thousand, so the answer is a bit less than three times eleven, which is 33. This is a typical heuristic process in that it is less demanding than a exact calculation and in that the answer it gives is not guaranteed to be accurate: in this case the correct answer is 31.7. It is also typical of the heuristics that we use in that its answers are close enough to be good enough for some purposes: the answer is four percent out in this case. Whether that is good enough would depend on the purpose of the calculation.

The distinction between heuristics and use of algorithms under processing limitations can be rather unclear. For example, in mental model theory, if people only generate one model in spontaneous reasoning with syllogisms, as Evans, Handley, Harper and Johnson-Laird (1999) propose[140], one can ask:

> Are people really applying an algorithm that constantly grinds to a premature halt, or are they applying the heuristic: *More often than not, the first possibility considered will enable a good approximation to the correct answer, and so no other possibilities ever need to be considered.* (Roberts, 2004, p. 239, his italics.)

Notwithstanding such borderline cases, the general distinction between heuristics and algorithms is clear enough, as I discuss in the next section. The

[140] This point was discussed in chapter 2. According to the theory, first a mental model is constructed from the input. Subsequently counterexamples consistent with the original information may be sought. Evans et al. say that in spontaneous inference this is the exception rather than the rule: "People can search for alternative models but do not necessarily do so spontaneously" (1999, p. 1507).

long-division example above illustrates an advantage of heuristics over most algorithms: in a short time, or subject to other processing limitations, one may not get any answer at all by use of an algorithm. The heuristics that we use are shortcuts that avoid this problem.

3.3.3 HISTORY AND USE OF THE TERM 'HEURISTIC'

The word 'heuristic' has been in English from around 1800, as an adjective for the first one-hundred and fifty years, and only in the second half of the twentieth century as a noun. The earliest citation in the OED is to Coleridge:

> *1821* COLERIDGE *Let. 8 Jan. (1971) V. 133, I am..getting regularly on with my* LOGIC *in 3 parts..3. Organic or Heuristic ()*

By 'heuristic logic' Coleridge meant the area of logic concerned with the rules governing discovery or invention. This sense of *heuristic* is still primary in some contemporary dictionaries, e.g. "1. enabling a person to discover or learn something for themselves." (Simpson & Weiner, 1991). By the early twentieth century, a connotation had accreted that heuristic meant *merely* useful in discovery and not certain: "Einstein used the term *heuristic* to indicate a view that was incomplete and unconfirmed, but nonetheless useful." (Marsh, Todd, & Gigerenzer, 2004, p. 274). Around this time, in early psychology, the word was used to describe "useful mental shortcuts, approximations, or rules of thumb used for guiding search and making decisions." (Marsh, Todd, & Gigerenzer, 2004, p. 274) From here it is a short step to the second sense in contemporary dictionaries: "[In] Computing[:] proceeding to a solution by trial and error or by rules that are only loosely defined." (Simpson & Weiner, 1991) This definition is confused – some heuristic procedures are precisely defined and do not involve trial and error – but it is at least clear that heuristics are in contrast with procedures that are guaranteed to find the correct answer, that is, algorithms. For this sense, and for the use as a noun, the OED cites two papers by Newell, Shaw and Simon (and see also Simon & Newell, 1958):

> *1957* A. NEWELL *et al. in Proc. Western Joint Computer Conf. XV. 223 A process that may solve a given problem, but offers no guarantees of doing so, is called a heuristic for that problem. Ibid,. For conciseness, we will use*

> *'heuristic' as a noun synonymous with 'heuristic process'. 1958 IBM Jrnl. Res. & Devel. II. 337/1 For the moment..we shall consider that a heuristic method (or a heuristic, to use the noun form) is a procedure that may lead us by a short cut to the goal we seek or it may lead us down a blind alley.*

In this thesis, the word heuristic is used strictly according to Newell et al's definitions. A heuristic is a procedure that, unlike an algorithm, is not guaranteed to reach the correct solution to a problem. Heuristics that are worth using are also shortcuts: they lead (often enough) to solutions faster or with less effort than algorithmic procedures. These two properties can come apart (as noted by Roberts (2004, p. 235)). Evidently, there could be non-algorithmic procedures that are not shortcuts, in that they require more resources than algorithms for a given problem, or in that they lead nowhere useful[141]. The ones that are of interest for a given problem are those which combine – in some proportion – better answers than blind guessing with less effort than algorithms.

Some heuristics are only barely better than guessing, but appear to be used because they lower effort considerably. One well-known example is the 'atmosphere strategy' in reasoning about syllogisms. This is the combination of two rules of thumb: (1) *if either of the premises contains the quantifier 'some' then the conclusion contains 'some'*; and (2) *if either of the premises contains a negation – 'no' or 'not' – then the conclusion contains a negation.* This leads to a correct conclusion for some pairs of premises, such as premises in the form: *Some of the As are Bs; None of the Bs are Cs,* where the atmosphere strategy correctly produces *Some of the As are not Cs.* The strategy yields more wrong than right answers, however: it gets only 23% of the 64 combinations correct (although all the conclusions it produces are compatible with the premises) (Roberts, 2004, p. 237 including note 3). Despite this very low accuracy, the strategy appears to be fairly common. For example, Gilhooly and colleagues found around 20% of participants apparently using this strategy (Gilhooly, Logie, Wetherick, & Wynn, 1993).

There are also shortcuts that are in a sense algorithmic and therefore not strictly heuristics. Roberts gives some examples. For solving syllogisms there

[141] A heuristic that has both demerits for many problems might be: *consult a fortuneteller.*

are procedures that are shortcuts relative to exhaustive search, but which are guaranteed to give the correct answer if they are applicable at all. One is the 'twosomes rule': *if both premises contain the quantifier 'some' then there is no valid syllogistic conclusion (except a restatement of one or both of the premises).* For example, given the pair of premises 'Some men are mortal' and 'Some penguins are mortal' nothing additional follows, as the rule says. Another is the two-negation rule: *if both premises contains 'not' or 'no' then there is no valid conclusion (distinct from the premises).* (Roberts, 2004, p. 239) This rule correctly predicts that from 'No men can fly' and 'Penguins cannot fly' nothing additional follows. These rules are heuristics in the original sense of discovery procedures, but they are not heuristics in the strictest interpretation of the more modern sense since the answers they give are guaranteed to be correct. If one is less strict, they might qualify as heuristics in the modern sense in that they do not guarantee a correct answer for all syllogism problems: each gives no answers at all to problems that do not match the conditions in its antecedent, and the combination of the two rules makes no prediction for certain pairs of syllogism premises.

A further example is the cancellation rule for solving compass-direction problems. In these problems, participants are asked to say where, relative to the original position, one would end up after taking a step north, a step east, a step north, a step west, and so on. The problem can be solved by keeping track of the position after each step, or by the computationally easier strategy of cancelling out north steps with south steps and east steps with west steps and adding up what remains on each axis. Interestingly, many participants reject the cancellation strategy as invalid: that is, they think it is a mere heuristic, when in fact it is algorithmic (Roberts & Newton, 2003).

Heuristics guide an agent towards solutions. In the context of theoretical reasoning this guidance may simply direct the operation of value-preserving rules, or it may direct making approximations that are non–value-preserving. For example, in attempting to derive a sequent in propositional logic, a heuristic which can be used to direct operations is: *if the conclusion is a conjunction try to derive the two conjuncts separately and then use and-introduction.* For problems in many fields, a fruitful heuristic is to work backwards from the end-state[142]. Heuristics that direct the making of approximations are necessarily specific to a domain or a type of task. In physics, such a heuristic is: *treat the tangent of a small angle as equal to the angle (in radians).* Such heuristics, in contrast to value-preserving rules, are rules of thumb mandating the use of approximations, so that their output does not follow with necessity from the input.

Useful heuristics reduce the amount of processing necessary, typically by reducing the amount of information that needs to be taken into account in reasoning. Heuristics typically pick out some key features of a task to work with, as we have seen: for example, the angle of elevation in ball-catching; the presence of two 'somes' or two negations in the rules of thumb for syllogisms. Since these key features depend on the task, or the task domain, or the format of the task, heuristics are specific to a particular domain, task or task format. This is the case for non-heuristic shortcuts too, as for example for the cancellation rule in compass-direction problems.

Despite this domain- or task-specificity, one cannot safely infer from the observation of content effects that heuristics are in operation rather than domain-general forms of reasoning such as mental logic and Johnson-Laird's mental models. Representation of input may depend on content or context, as may the task attempted, as discussed in section 3.2. Nonetheless, when agents operate faster, more accurately or more comfortably with tasks structured in a particular way, it is a good rule of thumb to investigate the possibility that a heuristic is in operation.

3.3.4 HEURISTICS AND TRIAL-AND-ERROR SEARCH

Because they may not deliver the correct answer, when heuristics are employed it is often as a component in trial-and-error search. One can have trial-and-error search without heuristics to guide it, and heuristics can be used in one-shot problem solving with no element of trial and error, but the combination of the two makes good sense. Simon's picture of human rationality is, at base, trial-and-error search constrained by heuristics:

[142] Some very general heuristics of this type are listed by Nickerson, including the following: *strive to understand the problem, analyse ends and means, make assumptions explicit, work backward,* and *simplify* (Nickerson, 2004, p. 422ff).

> ... human problem solving, from the most blundering to the most insightful, involves nothing more than varying mixtures of trial and error and selectivity. The selectivity derives from various rules of thumb, or heuristics, that suggest which paths should be tried first and which leads are promising. (Simon, 1962, pp. 472–473)

The use of trial-and-error search for some problems is inevitable. There are many problems for which algorithms are known but computationally unreasonable. We have seen that there is combinatorial explosion in consistency testing of sets of propositions by the truth-table method. Other famous examples include the travelling salesman problem and chess. In chess, the criteria for a good solution are known, but at most positions it is impossible in practice to calculate the optimal move, even for the most powerful computers[143](Todd & Gigerenzer, 2000, p. 730). There are other problems for the solution of which there is either no algorithm or none that terminates in a finite number of steps[144]. An example is proving sequents in predicate calculus or in higher order logics; a more complex example is abductive inference. There are also problems which have algorithms that are known, are computationally tractable, but nonetheless exceed human processing limitations, such as finding a winning strategy in draughts or solving syllogism problems. We do not play draughts by solving the game at each move – although computers can, and it is within human limitations to play noughts and crosses this way. Only a few tens of diagrams, models or schemas are needed to solve all syllogism problems definitively, but the evidence is that humans do not reason this way with syllogisms unless trained. (Roberts, 2004, pp. 234–236)

Trial-and-error search can be used for a problem if a good solution can be recognised when it is found. Then the way to solve a problem is to generate a

[143] Human chess competence is obviously complex, but it must be a combination of unconscious and conscious heuristics: it cannot be algorithmic. For most positions, even the most powerful computers cannot discover the optimal move with certainty. Computer chess programs, like humans, use a combination of memory and heuristic calculation. They have the advantages of precise recall of stored positions and fast calculation. Humans chess players have the advantages of better pattern recognition and intuition that certain moves are not worth considering.

[144] Sometimes guaranteed termination *in a finite number of steps* is taken as a defining criterion of the term *algorithm.* I take no view on this aspect of the definition.

solution and evaluate it, accepting or rejecting it accordingly. If the solution is rejected, and time allows, a new solution can be generated and evaluated, and so on, until a good enough solution is found or the search is given up[145]. In principle the generation of solutions could be random, but it is more efficient for search to be guided by heuristics and by indications of whether progress is being made:

> [in trial-and-error search] the trial and error is not completely random or blind; it is, in fact, rather highly selective. The new expressions that are obtained by transforming given ones are examined to see whether they represent progress toward the goal. Indications of progress spur further search in the same direction; lack of progress signals the abandonment of a line of search. (Simon, 1962, p. 472)

For trial-and-error search it is essential that there be a stopping rule. Stopping rules can be defined precisely by the problem, or they may be be defined more loosely. As an example of a problem that defines its own stopping rule precisely, consider the task of finding a value of x that satisfies an equation such as the one in (17).

(17) $17 = x^3 + 2x^2 - 2x + 2$

The equation that is to be solved provides a stopping rule for the problem. Search can be stopped with certainty if the trial value of x makes the equation true. Simon calls problems 'well structured' when "the goal tests are clear and easily applied, and when there is a well-defined set of generators for synthesizing potential solutions." (Simon, 1997, p. 128) Many problems, lacking these properties, are 'ill structured' (Simon, 1997, p. 128). One cannot be sure of finding an optimal solution to a problem where the goal test is not clear or not easily applied, or to be sure that an optimal solution has been found if in fact it has. For problems like this, there has to be a stopping rule that accepts solutions that are good enough. That is, it is necessary to have a stopping rule that

[145] The idea of trial-and-error search proceeding one solution at a time is of course something of an idealisation, at least as a description of cognitive processes. There may be some competition between trial solutions at each stage, for example.

satisfices. Satisficing plays a key role in theories of bounded rationality and I consider stopping rules and satisficing in some detail in section 3.3.5 below.

Turning to the second property of well-structured problems, the generator for potential solutions, consider the well-structured problem of finding a maximum of a curve (see figure 3) defined by an equation, such as (18).

(18) $y = x^3 + 2x^2 - 2x+2$

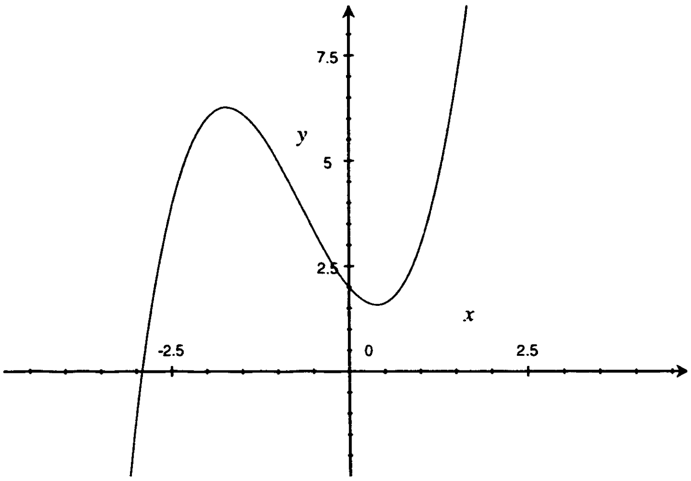

Figure 3: Cubic curve: $y = x^3 + 2x^2 - 2x+2$

Any maximum must lie on the curve: it must satisfy the equation. So one way to constrain solutions is to use the equation as a generator, generating a point on the curve, assessing whether it is a maximum, and if not, generating another point on the curve and assessing that, and so on until the solution (if there is one) is found. (In this case, there is a maximum at approximately $x=-1.7$, $y=6.2$.)

Abductive reasoning is much less well-structured, but there are similarities. Here too, to some extent the problem constrains what solutions are worth generating. In any solution to a problem of inference to the best explanation, the proposed explanation must logically support the fact that is being explained. Solutions with this property can be generated by a deductive device generating inferences from the input, as explained in chapter 2.

There is another way that the number of trials necessary in trial-and-error search is restricted. Systems for particular domains – agents, or component mental systems of the agent – accumulate expert knowledge of the types of problem that they encounter:

> The second source of selectivity in problem solving is previous experience. We see this particularly clearly when the problem to be solved is similar to one that has been solved before. Then, by simply trying again the paths that led to the earlier solution, or their analogues, trial-and-error search is greatly reduced or altogether eliminated. (Simon, 1962, p. 473)

There are two cognitive components behind expertise. One is the accumulation and refinement of heuristics that are useful in a particular domain or for a certain type of task. The second is the structuring of memory and cognition so that novel situations are recognised as similar to ones previously encountered and appropriate resources are automatically 'brought to mind' (i.e. activated). Both types of expertise narrow down the part of the problem space in which solutions are sought, decreasing the number of trials necessary to reach a solution, sometimes to the point that the first solution generated is usually correct, as Simon suggests.

According to this picture, expert behaviour in ill-structured problems is due to the same kind of problem-solving activity used with well-structured problems, but fuelled by recall of a large number of stored chunks of information:

> ... experts in any domain have stored in their memories a very large number of pieces of knowledge about that domain. Where it has been possible to measure the knowledge, at least crudely, it appears that the expert may have 50,000 or even 200,000 'chunks' (familiar units) of information – but probably not 5,000,000. (Simon, 1997, p. 128)

The 'chunks' of information are stored in such a way that features of a situation call to mind pieces of information with similar features, and which are therefore likely to be relevant:

> When the expert is confronted with a situation in his or her domain, various features or cues in the situation will attract attention. A chess player, for example, will notice such familiar cues as an 'open file', 'doubled pawns', or a 'pinned knight'. Each familiar feature that is noticed gives ac-

> cess to the chunks of information stored in memory that are relevant to that cue. (Simon, 1997, p. 128)

All (normal) humans are experts in many everyday problem domains, including the problem of utterance comprehension, in just this way: our memory is structured in chunks (so-called schemas or frames) which are recalled depending on features of the utterance and of the context. I expand on this treatment of utterance comprehension in chapter 5.

In this section, then, I have been agreeing with Simon (and with Sperber and Wilson, as I show in chapter 5) that there are a number of keys to understanding how limited beings solve 'ill-structured' problems: a) generation of trial solutions, followed by b) evaluation of each solution according to stopping rules, where c) the solutions generated are limited by heuristics that guide search, and d) the possibilities explored and the heuristics used are constrained by expert recall of relevant stored chunks of information on the basis of features of the problem. For problems that are not entirely ill-structured, the trial solutions generated may be limited by features of the problem and the stopping rule may make use of the problem definition.

Consider how real agents solve a very ill-structured problem that they face constantly: the problem of arriving at beliefs and/or intentions given the superabundance of information and potential inferential explosion. The central problem of rationality in the real world[146] is that there is a huge amount of potentially relevant information, and we must latch onto what is actually relevant and make use of it. There is a great deal of information in the input to our senses all the time, and we can seek out more information in long-term memory or in the external world. What should be done with all of this incident information? If all permissible inferences were performed on it there would obviously be an explosion of calculation. But some inferences must be performed, on pain of failing to recognise opportunities or dangers. So there are two problems at least: which information to entertain in working memory, and which inferences to perform on the contents of working memory.

Heuristic trial-and-error search cannot on its own be much help with this general problem, because the criteria for a good solution are not known: the

[146] i.e. bounded, adaptive rationality. The phrase is the subtitle of Gigerenzer, 2000.

problem is extremely ill structured. This problem must be dealt with by the way that cognition is set up, rather than by any one heuristic or stopping rule. Different processes, some heuristic, will be applied to the incoming information according to how fruitful their application is estimated to be: not by calculation, which would be a form of optimisation under constraints, but by the way previous experience (of the individual or the species) has set up the system.

3.3.5 SATISFICING

'Satisficing' is a term invented by Simon (Simon, 1956; Simon, 1957b; March & Simon, 1958)[147] and used by him in different but closely related ways. It is hard to pick these apart definitively, but I think that there are essentially two uses of this term in his work[148]. The broader use is to denote finding solutions that are good enough (i.e. useful, but not necessarily optimal). The narrower use of 'satisficing' denotes a specific kind of procedure by which satisficing in the broad sense might be carried out.

Satisficing procedures in the narrower sense are heuristic procedures of trial-and-error search where the stopping rule is based on an aspiration level regarding the object or solution (rather than a cue). They might be better referred to as sequential search with an aspiration-level stopping rule, since a) they can find solutions that are better than 'good enough' and b) satisficing in the broad sense of finding solutions that are good enough can be carried out in many different ways, and 'satisficing' in the narrow sense is only one of these. However *satisficing* is the term used in the literature, including the work of Gigerenzer and colleagues:

[147] The idea, but not the term, is present in Simon, 1955.

[148] Simon commented on his use of the term in a letter to Gigerenzer, reproduced in Gigerenzer, 2004, p. 406:

> ... I have used bounded rationality as the generic term, to refer to all of the limits that make a human being's problem spaces something quite different from the corresponding task environments: knowledge limits, computational limits, incomparability of component goals. I have used satisficing to refer to choice of "good enough" alternatives (perhaps defined by an aspiration level mechanism) or "best-so-far" alternatives to terminate selective search among alternatives—the latter usually not being given in advance, but generated sequentially. So one might apply "satisficing" to the "good-enough criterion" or to any heuristic search that uses such a criterion to make its choice.

> Satisficing takes the shortcut of setting an aspiration level and ending the search for alternatives as soon as one is found that exceeds the aspiration level, for instance leading an individual with Jack-Sprat-like preferences to marry the first potential mate encountered who is over a desired width. (Todd & Gigerenzer, 2000, p. 730)

Satisficing in this sense is in contrast with some other heuristics, in which search is stopped according to properties of the cue, rather than properties of the object selected. The 'take the best' heuristic (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 1999), for example, is used for tasks structured so that one chooses between two or more alternatives which are given and which are ranked according to a number of cues, also given. For example, one has to decide which of a pair of cities is larger, given such cues as whether or not a) each is a state capital, b) each has a premier-league football team and so on. Taking cues in order of subjective validity, the procedure stops with the first cue found that discriminates between the pair of objects. Therefore 'take the best' is a satisficing procedure in the broad sense but not the narrow one: when applied to a suitable problem it produces answers that while not guaranteed to be optimal are good enough for many purposes, but it does not do so by stopping search on the basis of the object found.

Satisficing in the narrow sense, sequential search with a stopping rule, is useful for choosing between alternatives which are not encountered all at once. This might be because the alternatives are spread out over space or time or both, as in house-hunting or job-hunting, or because the alternatives are solutions that are being generated sequentially, as in much product design, in some hypothesising and non-demonstrative inference, and (as I discuss in chapter 5) in utterance interpretation. In the first type of problem, the alternatives are discovered along the way and evaluated as they are discovered. In the second type of problem, the alternatives are not there to be found. They need to be generated, then evaluated. In both types of search, the set of alternatives is typically 'ill-bounded' (Simon, 1997, p. 126): there is no obvious natural end to the alternatives that might be considered.

The procedures employed are similar for the two types of problem. In search involving sequentially found alternatives from the environment, there must be procedures that guide the search, stopping rules to determine when the search should end (either because the most recent find is good enough, or because it seems unlikely that further search is worthwhile) and procedures for bringing appropriate information to bear on the decision to accept or reject an alternative. For example,

> [A job hunter] must not only have procedures for discovering prospective employers, but stop rules for determining when the search should end, and procedures for obtaining relevant information about each employment opportunity. (Simon, 1997, p. 126)

In sequential search involving generation of solutions, there must be also be heuristics that guide the search – in this case by guiding the generation of trial solutions; and as before, there must be stopping rules to determine when the search should end and procedures for bringing appropriate information to bear on the decision at each stage to accept and stop or reject and continue.

Satisficing in the narrow sense, then, is sequential search involving finding or generating alternatives, and evaluation of each alternative, which is either accepted, stopping search, or rejected, continuing search. The rule that stops search looks for a certain property of the object or solution being evaluated, and stops if the object or solution comes up to a certain standard. That is, there is a certain 'aspiration level' against which each object found or solution generated is compared. If the solution (or object) comes up to or exceeds the aspiration level then it is accepted and search is stopped. Unlike classical optimisation this kind of search does not involve consideration and ranking of all alternatives. While such search is called 'satisficing' it can do very well, depending on the suitability for the problem domain of the path followed and the aspiration level.

Various types of aspiration level are possible. The aspiration level may be set at the outset based on expectations about the domain: e.g. *buy the first coffee you find that is fair-trade or organic or* Illy-*brand*. Alternatively, the level can be set dynamically, during the search: for example by looking at candidates and setting an aspiration level. Given a set of alternatives that can only be accepted or rejected one-by-one (the example often used is potential mates), one form of strategy is to sample some proportion of the candidates, then choose the first encountered after that who (or which) is at least as good as the best encountered to that point. This is called 'best-so-far' search. In the problem of finding the best alternative from a sequence of known fixed length from an unknown distribution (rather unpleasantly known as the *secretary problem* or *dowry problem*), it has been shown that using an initial sample of 37% of the alternatives to set the aspiration level provides the highest likelihood of picking the best (Marsh, Todd, & Gigerenzer, 2004, p. 283).

A different form of dynamic aspiration level is one set according to the search cost so far. One such stopping rule is: *stop only when the payoff is worth the effort expended.* In environmental search this is usually an irrational strategy (but intuitively not uncommon: 'I've come this far, so I should press on'[149]) because the environment does not care how much effort you have put in – or to put it less anthropomorphically, mostly the environment is not structured so that the harder you have to look for an object, the more likely it is to be of high value[150].

In evaluation of sequentially generated solutions to a problem there is more to be said for a stopping rule like this. Each solution generated may build on previous solutions, as in a typical approach to design problems: in architecture, for example. Ideas from rejected solutions are refined and reworked in the production of a new trial solution, so there is some reason to expect the value of new solutions to correlate with the effort put into the process up to that point.

This may also be the case for utterance interpretation: each new prospective interpretation generated by the hearer's pragmatic faculties is likely to be a refinement of one that precedes it. But in the case of search for the intended interpretation of an utterance there is a special reason that it is rational to use a stopping rule of this type. In chapter 1, I mentioned Sperber and Wilson's postulate (1986) that in utterance interpretation there is a justified – although fallible – presumption that the correct interpretation will be worth the effort sunk into finding it. In fact, the presumption is stronger than that: the intended interpretation should be not only worth the effort the hearer has to put

[149] The 'fallacy of sunk costs' to economists. Shakespeare has Macbeth say: "I am in blood/ Stepp'd in so far that, should I wade no more,/ Returning were as tedious as go o'er" (Act III, scene IV).

[150] There may be exceptions. In competitive foraging, one might rationally expect to find more ripe fruit (for example) in locations that are hard to reach.

into finding it but also provides the best return for processing effort that the speaker had the ability and inclination to provide. The weaker presumption would justify a sequential search for the intended interpretation with a rule that search is to be stopped when the solution is at least worth the effort put in thus far. The stronger presumption makes it rational to use a different stopping rule: the solution must be at least worth the effort put in thus far *and* there must be no good prospect of a more valuable solution, up to the best the speaker was willing and able to provide. It is not clear whether a search with such a stopping rule should be regarded as satisficing in the narrow sense (albeit with a rather complicated aspiration level). It is certainly odd to call it that, because it typically does better than finding good enough solutions. It finds an optimal solution: the intended interpretation of an utterance, or something close to that. This is possible because it exploits a rather singular feature of its task domain.

As I also discuss in the next chapter, this feature, the presumption of optimal relevance, makes it rational to follow a least effort path for the problem of utterance interpretation, generating solutions in order of accessibility. As a general point, if (and only if) the solution lies on the least effort path, search that follows a least effort path is fast and frugal relative to other modes of search for the same problem.

A great deal of attention has recently been devoted to fast and frugal heuristics, particularly by Gigerenzer and colleagues (e.g. Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 1999; Gigerenzer & Todd, 1999). 'Take the best,' which I described above, is a fast and frugal heuristic, given that it consults as few cues as possible, in contrast to the classical norm for cue-based choice, Bayesian maximisation, which uses all cues, weighted by their respective validities.

As Todd and Gigerenzer comment, satisficing (i.e. sequential search with an aspiration-level stopping rule) and fast and frugal heuristics are "two overlapping but different categories of bounded rationality":

> there are some forms of satisficing that are fast and frugal, and others that are computationally unreasonable; and there are some fast and frugal

> heuristics that make satisficing sequential option decisions, and some that make simultaneous option choices (Todd & Gigerenzer, 2000, p. 731).

Satisficing (in the narrow sense) is already somewhat frugal when applied to an appropriate problem, since it "eliminates the need to compare a large number of possible candidates with each other, thus saving time and the need to acquire large amounts of information" (Marsh, Todd, & Gigerenzer, 2004, p. 283). An agent who uses sequential search with an aspiration-level stopping rule does not need to examine all the alternatives and need not try to work out what all the possible alternatives are (Simon, 1997, p. 119). This makes it possible to find solutions to open-ended search and decision problems that would otherwise be quite intractable. Other ways of looking at such problems have been previously discussed. In an unbounded approach it is as though all alternatives are known to the agent beforehand (and ranked in order of preference). In optimisation under constraints the solutions found are those that would be found by an agent who at each stage calculates the costs of further exploration of the problem space. Neither approach can be implemented computationally in a way that is frugal.

Sequential, aspiration-level search is more frugal than these alternatives, but not all such procedures are particularly frugal. To be frugal, a procedure must terminate quickly, which in most cases rules out random search of the problem space. 'Selectivity' in Simon's terms, in the form of problem recognition and guiding of search by appropriate heuristics, is vital for the frugality of sequential search since it reduces the number of trials necessary. Certain types of stopping rule would militate against frugality. The stopping rule must be computationally simple and must not require the gathering of too much information. In chapter 5 I assess the frugality of the comprehension procedure I have outlined above.

It has been implicit in this discussion of the overlap of sequential search and fast and frugal heuristics that any sequential search procedure with a realistic aspiration level is itself a heuristic, since it is not guaranteed to find an optimal solution. Such procedures also exemplify another point which I have left implicit until now: heuristics can be hierarchically arranged. A particular task in a particular domain may bring into use a certain heuristic, as for example (I argue) the task of interpretation of an ostensive act is accomplished by a procedure that follows a least-effort path and has the particular stopping rule described above. Within a sequential search procedure, the construction of each prospective solution may involve further heuristics, specific to features of the task: for example, one might learn to look for ironic interpretations of a particular speaker's utterances whenever that speaker avoids eye-contact while speaking.

3.3.6 HEURISTICS AND DEVELOPMENT

An obvious question to ask about any explanation in terms of mental processes is whether the explanation is developmentally realistic. Can we explain how an agent might come to have heuristics of the kinds that I have been suggesting? In the case of heuristics, what Chomsky calls Plato's Problem[151] is rather acute. Chomsky formulates Plato's Problem as follows (referring to Bertrand Russell's later work): "How comes it that human beings, whose contacts with the world are brief and personal and limited, are able to know as much as they do know?" (Chomsky, 1986, p. xxv; Chomsky, 1988, pp. 3–4)

It is particularly difficult to explain how we are able to develop heuristics for tasks we cannot solve algorithmically. If one does not know how to find the solution to a problem, how can one come up with a simple procedure that finds solutions[152] with the expenditure of little effort? Roberts makes this point (although without connecting it to the general philosophical problem):

> the difficulty in explaining the origin of many shortcuts is that it is hard to see how this process is constrained. If a person has difficulty in solving or understanding a problem, it is hard to see what criteria have been used to

[151] It is called Plato's problem because it is based on the question Plato raises in the Meno:

> Plato illustrated the problem with the first recorded psychological experiment (at least, a 'thought experiment'). In *The Meno* Socrates demonstrates that an untutored slave boy knows the principles of geometry by leading him through a series of questions, to the discovery of theorems of geometry. This experiment raises a problem that is still with us: How was the slave boy able to find truths of geometry without instruction or information?' (Chomsky, 1988, p. 4)

[152] This is rather close to Plato's own view of the problem. He has Socrates say "...if I don't know what something is, how could I know what that thing is like?" Meno: 71b (in Day, 1994). Nehemas defines Meno's paradox thus: "you can't look for what you don't know and don't need to look for what you know" (Nehemas, 1994, p. 227).

> generate a useful shortcut. Why not a useless one instead that results in worse than chance solution rates? (Roberts, 2004, p. 264)

The answer seems likely to lie in the same area as for much of the rest of cognitive science. Development is largely a matter of innately highly-constrained change in response to environmental triggers. It is probable that our evolutionary design equips us with fairly complete sets of heuristics for the basic accomplishment of certain vital tasks, including, for example, utterance interpretation, and also provides flexibility by allowing us to develop heuristics for other tasks and for sub-tasks, and perhaps to refine the innate heuristics. This flexibility would necessarily be constrained. Gigerenzer and colleagues' 'adaptive toolbox' (Gigerenzer & Todd, 1999; Gigerenzer & Selten, 2001) is one form that an explanation of constrained flexibility in development of heuristics could take. They suggest that heuristics are put together from simpler elements that are innately provided – the contents of the toolbox. The possibilities that are available, as in acquisition of concepts or of language, are those that can be constructed with the available building blocks, templates or parameters. This means that only certain heuristics will be possible.[153] Sequential search would presumably be part of the toolbox, as would aspiration levels that are set by prior expectations and can be adjusted during the search.

There is some support for the prediction that only certain types of heuristic are developed. In the compass directions task mentioned above, participants are occasionally found to use a *last-two* strategy. They ignore all the information up to the last two directions given, effectively assuming that all movement up to that point has cancelled out. This strategy seems irrational but has been found to do better than chance across a sample of typical task material – between 10% and 15% success versus about 3% for chance (Roberts & Newton, 2003). It is, therefore, a fast and frugal heuristic that is inaccurate but better than guessing, given its low cost. What is interesting is that of vari-

[153] Presumably other heuristics can sometimes be invented by systematic conscious effort, just as it is possible to invent concepts like *grue* (Goodman, 1954), and artificial languages with properties that do not conform to universal grammar. No natural language contains grue-type concepts: presumably humans do not form such concepts spontaneously (Sperber & Wilson, 1986, p. 69). Languages with properties that contravene UG are very difficult to learn (Smith & Tsimpli, 1995). Analogous heuristics should present similar difficulties.

ous other logical possibilities, including a *first-two* strategy and a *first-and-last* strategy, none is attested (Roberts, 2004, p. 264). It may be relevant that these strategies would require the use of long-term memory and would therefore involve greater effort. Perhaps the *last-two* strategy is due mainly to using whatever is left in short-term memory at the end of the task. As a general point, the heuristics that are feasible will be ones that work within the limitations of human memory structure and processing capacity.

3.3.7 RATIONALITY AND ADAPTIVITY

To conclude this chapter – and the part of this thesis which deals with general questions about rationality – I comment on a question that may have occurred to the reader. In chapter 2 I defined rationality as the possession of reasoning ability, and reasoning ability as consisting essentially in the ability to make value-preserving transitions. In the present chapter, I have been arguing that rationality is bounded and I also have said that it is rational to take shortcuts which lead to answers that are good enough, given the 'finitary predicament' of humans. There is an apparent tension between the idea that rationality is at its core an ability to make value-preserving transitions, and the idea that rationality is largely implemented through heuristics. In particular, it might be thought that in claiming that it is rational to take shortcuts because they make good enough decisions with necessarily limited resources, there is a commitment that rationality reduces to evolutionary adaptivity. Another way of seeing this question is as a clash between two views in which cognition is computation over mental representations: Fodor's view according to which such computation must preserve semantic value, and Simon's (and Gigerenzer's) view in which this is not of such importance as the solution of problems by simple means.

It is one of the main contentions of this thesis that this apparent problem is solvable, and that there is no contradiction between a traditional view that rationality is the possession of reasoning ability, which is essentially the ability to make transitions that preserve rational acceptability; and a view that real rational beings must make decisions that exploit features of the environment to enable them to work within their limited cognitive means. A second major contention of the thesis is that resolution of this apparent problem is a necessary starting point for a theory of pragmatics, since pragmatics involves finding interpretations that are logically warranted by utterances, and finding them fast.

These two main points are reflected in two subsidiary contentions which both amount to exegesis of Grice's work. I have argued that Grice was aware that much reasoning involved shortcuts "made possible by habituation and intention", and was not thereby persuaded to abandon his claim that the core of rationality is the ability to make steps that preserve value. In chapter 4, I argue that the strong parallel between Grice's views about reasoning and about the role of rational reconstruction in utterance interpretation means that he was implicitly committed to the position that utterance interpretation was a form of reasoning, even when the hearer is not conscious of constructing a truth-preserving argument.

One manifestation of the problem I am discussing here is that, perhaps contrary to common sense, it is not the case that if one knows rules that will lead with certainty to the solution of a problem, rationality demands that one use those rules[154]. This is an intuitively attractive position, but one that has to be rejected if the pervasive use of heuristics can be rational.

Heuristics, in the narrow sense of procedures that do not guarantee correct answers, are often used in cases where algorithms exist, in preference to those rules. Can their use be rational? It can, for two reasons. The first reason is that a heuristic may require much less time and effort than a full algorithmic derivation of an answer. Having some answer, and having it quickly and with little effort, can be preferable to having no answer for a long time while an algorithm is used to work out an answer, at relatively much greater effort. The second reason is that a heuristic can be very accurate when applied to a problem with the right environmental structure. The ball-catching heuristic from the first chapter is a good illustration of both points.

So rationality does not demand that whenever we must solve a problem for which an algorithm is known we use it. It is not rational to try to play chess by calculating the optimal move, because the universe would end before that method would yield an answer. It is not rational to try to play draughts by calculating the optimal move, because, unless one's life depends on winning, it

[154] See Brown, 1988 for this intuition, which he also rejects, for different reasons.

is not worth so much more effort than using less demanding heuristics. We have to make inferences and decisions fast enough or it will be too late, and we have to make them at limited cost because there are so many demands on our resources. Moreover there are no algorithms for certain key types of problem, such as inductive inference. In inference to the best explanation, one could follow truth-preserving inference rules forever with no guarantee that one would find an explanation for the observation that one started with.

We cannot do without heuristics, therefore. One might then ask: can we make do without the ability to make value-preserving transitions? The answer is that we cannot. We have to be able to make valid inferences because they tell us how things are: if some proposition p is true and another of the form $p \rightarrow q$ is true, then the proposition q is also true. The ability to make value-preserving inferences is simply the ability to recognise such things, to work out what follows. This ability may be put to use in the service of either algorithmic or heuristic procedures. Above I argued that we cannot stamp inconsistency out totally from our set of beliefs, because there is no computationally reasonable way of doing so. Similarly we could not hope to reach all valid conclusions that are logically supported by our set of beliefs in long-term memory in a lifetime, even setting aside the issue of trivial consequences. However, I am not arguing that failure to recognise inconsistency or failure to draw inferences are harmless. On the contrary, failures of this sort can be serious. Cherniak gives an example:

> Smith believes an open flame can ignite gasoline..., and Smith believes the match he now holds has an open flame..., and Smith is not suicidal. Yet Smith decides to see whether a gasoline tank is empty by looking inside while holding the match nearby for illumination. (Cherniak, 1986, p. 57)

Cherniak hypotheses that what happens is connected with the structure of memory and retrieval from long term storage. Smith's belief that a match is a means of illumination is active, but this does not result in a check against the category *means of ignition* since his goal at the time is illumination. So the two relevant beliefs about matches are not both in short-term memory simultaneously, and the crucial conclusion is not drawn. I would add that the problem in such cases is that search stopped too early, that is, that processing was in-appropriately shallow. The cognitive mechanisms looking for relevance have failed, since there was a highly relevant conclusion to be derived from a slightly longer search and reasoning process.

Cherniak's point is that given that we are finite beings, memory must be structured, and this means that it is inevitable that we will sometimes make mistakes when we do not recall information that would have been highly relevant. But this example can also be seen as demonstrating why we need the ability to make truth-preserving transitions. If the relevant beliefs were recalled to short-term memory and the conclusion was not drawn, this would be a more serious failure of rationality. If there were a creature that never drew logical conclusions from beliefs in its short-term memory, then it would not be rational at all (this is Cherniak's minimal inference condition (1986, p. 57)).

To recapitulate, my claim is there is no clash between the views of rationality presented in chapter 2 and chapter 3. Both aspects are necessary for a full theory of human rationality. We have to be able to make inferences or we are not rational at all. On the other hand, we have to make inferences fast enough or it will be too late. This is a non-trivial problem given that we are very limited beings and that abductive inferences are of particular importance in understanding the world around us. The superabundance of information compounds the problem further.

I said above that there is a question about whether the rationality of a creature or a procedure reduces to how well adapted it is. The position I have adopted (in chapter 2) about the definition of rationality is that we are rational beings in that we have some ability to perceive or discover logical relationships, in particular, logical consequences of our beliefs, by constructing chains of inferences. Creatures that can do this are rational. Creatures that cannot are a-rational or non-rational.

One of the things that our reasoning abilities allow us to see is that creatures have to behave in certain ways if they want to survive and prosper. We can see that creatures have to make good enough choices fast enough, or they will not tend to do well. We can see (or work out) that this applies just as much to creatures which have the higher-level abilities involved in rationality: abilities to manipulate propositional representations according to their forms.

From this perspective one can agree with Gigerenzer and colleagues that "the ultimate test of the 'rationality' of a heuristic can be found in its fitness consequences relative to real constraints and real environmental structures." (Marsh, Todd, & Gigerenzer, 2004, p. 275), including heuristics for reasoning.

That is, we can call adaptive behaviour rational; and we can also call the heuristics or other processes or faculties that generate it rational. It is important to bear in mind that this is hypothetical rationality of this form: *If you want to survive and prosper, you should behave like this*; or *If you want to survive and prosper, it makes sense to have the abilities required to behave like this.* There is nothing wrong with talk of the rationality of component systems of creatures, including decision-making systems, from the perspective of an imagined designer (Grice's 'genitor' (1974)). It is rationally acceptable that a creature be equipped with such and such a capability, from the genitor's point of view, and this applies just as much to equipment for flying or navigating by sonar as to reasoning.

However it would be a confusion to say that creatures that are well-adapted are thereby rational. It would also be a confusion to say that the faculties that contribute to a well-adapted creature's adaptedness are rational, except in this hypothetical sense: it is rationally acceptable that they have such faculties, given that they face the problem of survival.

The message of chapters 2 and 3 of this thesis is that if we want to understand an inferential reasoning process within cognitive science, it is inevitable that we bring together ecologically plausible assumptions about rationality with the kind of logical-causal picture needed for a psychologically realistic version of reasoning and inference. It is rational (from the genitor's point of view) that a creature that is rational (in the straightforward sense that it can reason) should be able to reason fast enough and well-enough that its reasoning abilities are useful to it.

In the next chapter I return to pragmatics, applying my view of rationality to an inferential-intentional theory of utterance processing.

Chapter 4 · Reasons, reasoning and meaning

4.1 INTRODUCTION

In chapters 2 and 3 I have set out my view of rationality. I have claimed that to be rational (at least in the theoretical realm) is, essentially, to possess reasoning ability, where reasoning ability is the ability to make transitions that preserve a certain type of acceptability. In theoretical reasoning such transitions preserve truth, or at least warrant. A second aspect of rationality follows from the demand that explanations in science be explanatory. I have argued that this means, first, that central cognition, including reasoning processes, should be seen as computations over representations, and secondly that these computations must be tractable, and in many cases fast and frugal. I have tried to show in general terms how these criteria can be met. In the present chapter I move from the general to the specific, attempting to cast in this mould the pragmatic processes involved in utterance production and, particularly, utterance interpretation.

In chapter one I set out the fundamentals of a broadly Gricean view of pragmatics, according to which making an utterance involves providing evidence to the hearer of certain intentions, on the assumption that the hearer is rational and will be able to infer the intentions from the evidence. This means that interpreting an utterance is a matter of attempting to infer a speaker's meaning, that is, certain of the speaker's intentions, from the evidence provided. This is a form of inference to the best explanation. I discussed some alternative views according to which either inference or the speaker's intentions, or both, need not be involved. While I do not return to those alternative views here, the business of this chapter can be seen as the construction of a riposte to one of their motivations. As I showed in chapter 1, one motivation for such views is a concern that a Gricean picture of utterance interpretation is psychologically unrealistic. Gricean pragmatics relies for its explanatory force on quite elaborate schemas for the hearer's inference about the speaker's intentions. The worry raised by theorists such as Millikan is that such schemas are just rational reconstruction with nothing to do with the processes actually involved in utterance interpretation.

It should be clear that my intention in exploring Grice's distinction between the 'hard way' and the 'quick way' of reasoning in chapter 2 was to establish that a theory of reasoning as purposeful value-preserving transitions is compatible with the claim that in much actual reasoning shortcuts are used. In the third section of this chapter I look at the Gricean picture of pragmatics in more detail. I claim that utterance interpretation is reasoning, regardless of whether it is carried out the hard way or the quick way. I also maintain that Grice was implicitly committed to this position, given the similarity between what he says about calculability in pragmatic inference and about what counts as reasoning.

Most pragmatic inference is carried out quickly, and largely unconsciously. However, it seems that the inference can be reconstructed after the fact, if the input (the utterance and context) and output (the interpretation reached) are known. François Recanati (2002b; 2004) has argued that the conscious availability to the inferrer of pragmatic inferences (and their inputs and outputs) is essentially connected with their status as inferences, or as cases of reasoning. I argue that this is not correct, and reject the distinction he finds on this basis between associative, 'primary' and inferential, 'secondary' pragmatic processes. It is my position, first, that all of speaker's meaning must be grasped inferentially, notwithstanding the Gricean point that implicatures are inferred from what is said (in some sense of the phrase), and, secondly, that it is better not to put too much stock in whether a particular process is conscious or not when trying to understand its properties.

In the fifth chapter, I look in more detail at how utterance interpretation is achieved in a computationally tractable, psychologically realistic way. The broadly Gricean view of pragmatics that I take defines the problem that the utterance interpretation system must solve. It shows what kind of inferences the pragmatic inference system must achieve. However, as mentioned in chapter 1, I do not adopt the specifics of Grice's explanation of how implicatures are derived, the cooperative principle and maxims. Communication is not a cooperative, but a coordinative endeavour; and Grice's system of implicatures arising from both violations and non-violations of maxims does not appear promising as a computationally tractable, psychologically realistic account.

In section 4.2, effectively a prologue to the current chapter, I look at an objection to the study of the use of language, particularly utterance production. According to this point of view, which is sometimes attributed to Chomsky, there is no way to scientifically study, in terms of mental causation, problems that involve free choice guided by reasons. I suggest the division of utterance production into two components, one amenable to study, one apparently not.

4.2 DESCARTES' PROBLEM AND PRAGMATICS

4.2.1 CHALLENGES TO A REASONS-BASED VIEW OF UTTERANCES

> As soon as questions of will or decision or reason or choice of action arise, human science is at a loss. Noam Chomsky, in a television interview.[155]

The idea of scientifically investigating the processes involved in production of utterances and in interpretations of utterances might be challenged from two perspectives at least. One line of argument, exemplified in Ryle, 1949, is that talk about mental entities is superfluous or meaningless. This point of view has been mentioned and countered in the introduction and in chapter 2 since it provides a preemptive challenge both to cognitive science as a whole and to the idea of treating central cognition and reasoning in terms of computations over mental representations.[156]

A second objection has been attributed to Chomsky. This is the idea that language use does not form a suitable domain for scientific enquiry, since it

[155] Chomsky has made similar remarks in published work. Compare, for example, "The traditional issues of will and choice remain off the agenda of empirical enquiry." (Chomsky, 2003, p. 262)

[156] Ryle would have regarded these approaches as mired in what he called the 'intellectualist legend' (Ryle, 1949, p. 29)

involves poorly understood questions about human creativity, free will and agency.

4.2.2 CHOMSKY AND PRAGMATICS

As noted in the introduction, Chomsky is one of the fathers of modern cognitive science, famous for his opposition to views such as Ryle's that outlaw talk of unseen mental events or states. In Chomsky's view, modern cognitive science takes off from the 'second cognitive revolution' of the mid-twentieth century, which "is concerned with states of the mind/brain and how they enter into behavior" (Chomsky, 1991a, p. 5). On the face of it, the study of the inferential component of communication in terms of a computational theory of mind is a natural application of this approach.

However the view has been attributed to Chomsky that an explanation of language use and the understanding of utterances in terms of "states of the mind/brain and how they enter into behavior" is problematic and perhaps impossible. The attribution is made (by, for example, Carston (2000), McGilvray (2005) and the present author (2005) among others) because of remarks that Chomsky has made about the problem of free will and creativity, or 'Descartes's problem' as he calls it. Gricean pragmatics, at least in psychologically realistic forms, involves study of how a speaker's reasons are causally effective in production of an utterance. Some of Chomsky's remarks strongly suggest that he regards it as impossible to study the causes of intentional behaviour, including language use. However, Kasher has argued that this is a misinterpretation of Chomsky's views on pragmatics. The problem of language use, he says, can be divided into two parts, one of which can be approached scientifically. I agree with this view of the substantive question, reserving judgement on the correct interpretation of Chomsky's remarks.

Given a conception of having or knowing a language as a cognitive state, Chomsky says that three fundamental questions arise (Chomsky, 1991a, p. 6). The first is Humboldt's problem: "what constitutes knowledge of language?" (Chomsky, 1991a, pp. 6, 7) The second is the question of how such knowledge is acquired. This is an instance of Plato's problem (Chomsky, 1991a, p. 15), which was discussed in chapter 3, above, in relation to the acquisition of heuristics. The third question is how such knowledge is put to use. This is Descartes' problem (Chomsky, 1991a, pp. 17–20).

Chomsky uses the term *Descartes' problem* for the problem of explaining the use of language because of Descartes' view[157] that, "normal human speech is unbounded, free of stimulus control, coherent and appropriate, evoking thoughts that the listener might have expressed in the same way – what we might call 'the creative aspect of language use.'" (Chomsky, 1991b, p. 40)[158] The argument, which goes back to Chomsky's dismissal of behaviourism in his review of Skinner (Chomsky, 1959) depends on the observation that what a language user might say is not predictable from the circumstances she is in.

This problem is one aspect of a "general problem concerning human action" (Chomsky, 1991b, p. 40) that arises if we seek to understand human action in terms of computations over representations. If you know the internal state of a computational system, an automaton in Descartes' terms, and its environmental inputs, then you know what it will do, because its behaviour is a function of its state, perhaps acting on information from the environment. Chomsky takes the view that humans differ from computational systems in this respect because they are only "incited and inclined" towards certain actions, not "compelled" to perform them. (Chomsky, 1991b, p. 40). Thus truly creative activity, of the sort that springs from and exemplifies free will, may be "beyond the powers of any automaton" (Chomsky, 1991b, p. 40). Questions of human creativity may therefore lie beyond human investigation: in Chomsky's terms they may be mysteries rather than problems, where problems are "questions that we seem to be able to formulate in ways that allow us to proceed with serious inquiry and possibly to attain a degree of understanding"

[157] Descartes' view was couched as a thought experiment about the capabilities of perfect automata. They "could never use words or other signs arranged in such a manner as is competent to us in order to declare our thoughts to others: for we may easily conceive a machine to be so constructed that it emits vocables, and even that it emits some correspondent to the action upon it of external objects which cause a change in its organs ... but not that it should arrange them variously so as appropriately to reply to what is said in its presence" (Descartes, 1912, pp. 44–5, Discourse V). This led him to conclude, in the terms of his dualist ontology, that "the faculty responsible for language must be a faculty of an immaterial substance – matter could not account for the infinite flexibility and creativity manifest in language use" (Antony & Hornstein, 2003).

[158] On the creative aspect of language use, see also Chomsky, 1964; Chomsky, 1966; Chomsky, 1986; Chomsky, 1988.

(Chomsky, 1991b, p. 41) while mysteries are "questions that seem to elude our grasp, perhaps because we are as ill-equipped to deal with them as a rat is with a prime number maze" (Chomsky, 1991b, p. 41).

A different way of making essentially the same point rests on the assumption that the study of language use requires some scientific understanding of human intentions. Chomsky's view is that "General issues of intentionality, including those of language use, cannot reasonably be assumed to fall within naturalistic inquiry" (Chomsky, 1995, p. 27). Remarks of this sort have led, naturally enough, to the interpretation that I mentioned above – that Chomsky may hold the view that scientific pragmatics is impossible:

> Chomsky and at least some other generativists are sceptical about the feasibility of pragmatics, where pragmatics is conceived of as an account of utterance interpretation. Such a pragmatics is generally taken to involve the inferential recognition of speaker's intentions ... and for Chomsky, matters involving human intentions may well lie beyond the scope of scientific enquiry (Carston, 2000, p. 28; citing Chomsky, 1995)

However Kasher rejects this interpretation of Chomsky's views, for reasons I return to later in this section:

> A couple of times recently, we have heard the view that pragmatics is impossible being ascribed to Chomsky, on grounds of his attitudes towards 'Descartes's Problem.' ... this is a misguided ascription, resting on a confusion of 'Descartes's Problem' with the pragmatic problem. (Kasher, 1991, pp. 143, note 16)[159]

In his exposition of Chomsky's views on creativity and language, McGilvray makes clear that in his view the claim is that it is the creative, unbounded nature of language production that makes it an unsuitable subject for scientific study. As an example, Gertrude, during a conversation about computer chips, might suddenly say:

[159] See the main text of (Kasher, 1991) for discussion and references relating to Chomsky's remarks on language use, especially pp. 123-4.

(19) I'm going to join the Canadian bobsled team. (McGilvray, 2005, p. 221)

As McGilvray says, "Her environment does not cause the sentence. She need not say anything at all, and could have said any number of things." Granting this, it is not clear that we must reach the conclusion that language use cannot be fruitfully studied. The key problem with that contention is that although language use is unbounded and not caused by input from the environment alone, it is typically "appropriate and coherent to circumstances", as McGilvray notes (2005, p. 221). Speakers generally have reasons for what they say, and in my opinion these reasons have causal efficacy (although they may very well not have reflected consciously on those reasons), as I discuss in section 4.3. As McGilvray explains:

> Perhaps [Gertrude] is letting her companions know she is bored and wants to talk about something else, or reminding them that their meals are getting cold. Perhaps she really wants to join a bobsled team. So while circumstances do not cause her sentence, it is appropriate to them: she has a reason – perhaps several – to say what she does. (McGilvray, 2005, p. 221)

Pragmatics attempts the systematic study of the way that a speaker's reasons and purposes lead to her saying what she does in the way that she does, and the way that hearers work these reasons out from their product, utterances. Of course it is possible that while speakers have reasons for their utterances, those reasons and any causal role they play are not suitable for scientific study. McGilvray's remarks echo comments that Chomsky has made which suggest that this may be his view, for example:

> Human action is coherent and appropriate, but appropriateness to situations must be sharply distinguished from the causal effect of situations and internal states. There is little reason to suppose that human behaviour is caused, in any sense of the word we understand. (Chomsky, 1991b, p. 41)

The position on language use that has been attributed to Chomsky is that, although speakers have reasons for their uses of language, we can only usefully study the language system that they use, *not* the 'production problem': the reasons why they use it in a particular way and how their uses of language come to be coherent and appropriate. Indeed, Chomsky believes that the production problem may be a mystery for humans (Chomsky, 1991b, p. 41).

In my view, it is clear that as tasks for the mind/brain, the speaker's task is not symmetrical with a hearer's comprehension of a speaker's use of language. It is not the case that speaker and hearer are just doing the same things, but in the opposite order. While a speaker is, in a sense, free to say anything at all (or to make no utterance), a hearer has a much less open-ended task. The hearer's task is to assign an interpretation – which must be near enough to what the speaker intended – to an utterance once it is made (by inferring what the speaker meant by her utterance, in Gricean theories).

Therefore one might suppose that Chomsky's sceptical remarks about the study of language use are meant to apply only to the speaker's creative task rather than the more constrained task of the hearer. There is some reason, though, to suppose that Chomsky believes that scientific investigation of the way a hearer arrives at an interpretation is also hopeless. He says that language competence and parsing can be studied, but that:

> There is also a further problem, which we can formulate in vague terms but which cannot be studied in practice: namely to construct an 'interpreter' which includes the parser as a component, along with all other capacities of the mind – whatever they may be – and accepts non-linguistic as well as linguistic input. This interpreter, presented with an utterance and a situation, assigns some interpretation to what is being said by a person in this situation. The study of communication in the actual world of experience is the study of the interpreter, but this is not a topic of empirical enquiry for the usual reasons: there is no such topic as the study of everything. ... The proper conclusion is not that we must abandon concepts of language that can be productively studied, but that the topic of successful communication in the actual world of experience is far too complex and obscure to merit attention in empirical inquiry. (Chomsky, 1992, p. 120)

According to this point of view, the problem of 'perception'[160], which "is concerned with the process by which a person assigns a structural description to a presented expression in a particular situation"[161] (Chomsky, 1991a, p. 18) subdivides into the study of the parser and the study of the full interpreter. The motivation for this subdivision is seeking "to isolate elements of the problem that can be subjected to inquiry, under appropriate idealisations, their appropriateness determined, as always, by the explanatory success achieved by using them." (Chomsky, 1991a, p. 6) On this view, the parser is "a feasible subject of inquiry" (Chomsky, 1991b, p. 40) (although in Chomsky's opinion the concept is not as clear as might be wished; certainly not as clear as the concept of linguistic competence). The full interpreter on the other hand "may not be a feasible subject of inquiry... : it is a too-many-factor problem" (Chomsky, 1991b, p. 40)

I claim that we can study not only the hearer's inferences about the speaker's intended meaning, but also the reasons that the speaker has, given a meaning that she wants to convey to the hearer, for making one utterance rather than another. This point was, of course, made by Grice in his work on meaning.

I would like to suggest that the 'problem of language use' should be broken into two parts. There is the question of what a speaker might want to communicate in a particular situation. Here I agree with Chomsky (and Descartes) that this question is not amenable to scientific study in terms of properties of automata, or the modern equivalent, computations over mental

[160] I do not regard arriving at speaker's meaning as part of perception. That is because it has an inferential component. One does not perceive what the speaker means, one works it out. (It is not relevant that in ordinary use one talks of 'seeing' a speaker's meaning, since we talk this way about inferences generally: e.g. *Now I see how he did it - he introduced the snake through the window.*)

[161] In fact it is not just a structural description of the presented expression we are after but also of the explicit meaning it expresses and the implicatures it carries in the context. On an inferential view, the structural description from parsing must be part of the input into the process which infers speaker's meaning.

representations (at present, at least),[162] perhaps because it is bound up with questions about free will.

There is also the group of questions about a speaker's reasons for producing a particular utterance to convey a given intended meaning in a certain situation, and the inferences a hearer will make about intended meaning, given an utterance. This second group of questions, on the face of it, is much more approachable. The speaker's aim is given. What is to be investigated is how she chooses means that are rationally appropriate to achieve it. The hearer's interpretative task is also, and perhaps more obviously, an exercise of rationality.

A full-blown scepticism about the study of language use would rely on not seeing the two groups of questions as separable, or perhaps on thinking that the second group is no more approachable than the first. In my opinion, Grice's work on meaning can be seen as identifying this second group of questions and showing how they might be made tractable in terms of general considerations about rationality.

Kasher had made essentially this point (1991) some years before I made this suggestion (2005)[163]:

> The more we explain pragmatic facts in terms of a general intentional action theory as applied to instances of language use, the closer we come to solving parts of 'Descartes's problem'. Creative use of language can be factored into (a) creative choice of ends and (b) rational pursuit of those ends. Factor (b), of the rational pursuit of given ends, seems to be amenable to explanations in terms of general rationality principles, which are parts of a general intentional action theory. However factor (a), of the cre-

[162] Questions of this kind may be amenable to scientific investigation in statistical terms, as for example in the social psychology literature where it has been found that certain stimuli prime actions – they make certain types of behaviour more likely. (Bargh, 2006, is a recent survey of priming of representations and behaviour.)

[163] I was not aware of this paper by Kasher at that time. I would like to take this opportunity to acknowledge his priority on this point.

Kasher's views on this issue are expressed somewhat differently from mine since he advocates a competence theory of pragmatics. I agree with him that what pragmatics studies in utterance production is "the conditions which constrain 'what we say' in a context" rather than "an understanding of the creative aspects of 'what we say and why we say it" (Kasher, 1991, p. 127), but not that it is the study of the knowledge of those conditions.

> ative choice of ends, does perhaps constitute an unsolvable problem, a mystery." (Kasher, 1991, p. 141)

Of course, any view that language use is unlikely to yield to naturalistic enquiry faces the problem that pragmatics appears to be a successful scientific research programme judging by the usual standards. Among other merits, it offers unified explanations of phenomena previously thought unconnected; it inspires experimental work; and its conceptual foundations cohere with those of other branches of cognitive science. In the next section I look at the Gricean foundations of modern pragmatics in some detail, with particular stress on how considerations of rationality are central to the picture.

4.3 GRICE, REASONING AND PRAGMATICS

> the use of language is one among a range of forms of rational activity (Grice, 1989b, p. 341)

In the Retrospective Epilogue to a selection of his papers (Grice, 1989c), Grice picks out eight 'strands' from his philosophical writings. This section (and indeed, this thesis as a whole) readdresses part of his Strand Six: "the idea that the use of language is one among a range of forms of rational activity and that those rational activities which do not involve the use of language are in various ways importantly parallel to those which do" (Grice, 1989b, p. 341), which flows from his work on meaning (Strands Four and Five).

4.3.1 GRICE, PRAGMATICS AND EXPLANATION

One of Grice's contributions to pragmatics was to focus attention on its connections with rationality, inference and reasoning. He suggested that talking might be seen "as a special case or variety of purposive, indeed rational, behaviour" (Grice, 1975, p. 47) and that aspects of a speaker's meaning which go beyond sentence meaning must be inferred (in effect making the point that they cannot be decoded, as I discussed in chapter 1).

For Grice, communication involves reasoning in at least two ways. The first connection is that communication of implicatures depends on rational interaction[164] between speaker and hearer, and implicatures must be derivable by a reasoning process. Grice called this property *calculability*[165]. Secondly, Grice earlier argued that the analysis of speaker meaning more generally involves an appeal to reasons. For something to count as the meaning of an utterance, it must be a response that the hearer has to the utterance because of the intention behind the utterance, both in the sense that the recognition of the intention causes the response and in the sense that it provides a reason to have that response. Thus the link between the recognition of the speaker's intention and the hearer's interpretation of the utterance is a causal process that preserves rational acceptability: a reasoning process as defined in chapter 2. I discuss each of these connections in greater detail below. The key point is that reasons and reasoning are important foundations of Grice's work on communication and meaning, just as they are central to a great deal of Grice's philosophy.

Grice was committed to understanding humans as rational agents, that is, as beings who have reasons for their actions and attitudes. This meant that he would try to understand actions and attitudes partly in terms of the reasons people might (or should) give for them and the reasoning they might (or should) follow to work out which attitude to adopt or action to take. Richard Warner identifies this as "a key feature of Grice's philosophical methodology":

> Given the task of providing a philosophical account of some kind of attitude or action, or some other psychological aspect of life (for example, intending to catch the 5.01 train, doing one's duty, living a happy life), Grice would ask, "How would a person explicitly reason his way to that attitude, action, or realization of that aspect in his or her life?" (Warner, 2001, p. x)

This way of proceeding is exemplified in Grice's work on communication. The Cooperative Principle and conversational maxims can be seen as Grice's answer to a question he posed for himself: supposing that people are rational agents, how should one expect them to behave in conversation and other situ-

[164] Specifically, for Grice, cooperation: see chapter 1.

[165] It is only conversational implicatures that must be calculable, for Grice. I do not discuss conventional implicatures.

ations in which they have the goal of communicating? His conjecture is that they would cooperate, to some extent, and their communicative behaviour would be governed by certain rules or principles:

> I would like to be able to show that observance of the Cooperative Principle and maxims is reasonable (rational) along the following lines: that anyone who cares about the goals that are central to conversation/communication ... must be expected to have an interest, given suitable circumstances, in participation in talk exchanges that will be profitable only on the assumption that they are conducted in general accordance with the Cooperative Principle and maxims. Whether any such conclusion can be reached, I am uncertain. (Grice, 1975, p. 49)

My view is that Grice's Cooperative Principle and maxims are not compatible with reasonable assumptions about rationality and the communicative situation. I have given some reasons in chapter 1 to think that communication is coordinative rather than cooperative in Grice's sense. I make some mention below of the shortcomings of his system of maxims. In my opinion it is the other connections between a Gricean picture of reasoning and Grice's work on communication and meaning that form the foundations of inferential theories of communication.[166]

As Sperber and Wilson say, Grice's analysis of meaning could be used as the starting point for a theory of meaning or "as the point of departure for an inferential model of communication" (Sperber & Wilson, 1986, p. 21). Taking it the second way, and exploring the role of inference, reasons and reasoning in Grice's account of the way speaker's meaning is arrived at, the intimate links between Grice's work on meaning and his work on conversation and implicature become clear. On the relation between Grice's work on meaning and his work on a theory of conversation, I can do no better than to agree with Stephen Neale, who writes that:

[166] As Grice says of his Strand Six, the thesis that language is a rational activity:

> This thesis may take the more specific form of holding that the kind of rational activity that the use of language involves is a form of rational cooperation; the merits of this more specific idea would of course be independent of the larger idea under which it falls. (Grice, 1989b, p. 341)

> There is no doubt that Grice's work on language and meaning constitutes a more powerful and interesting contribution to philosophy and linguistics when it is not seen as comprising two utterly distinct theories. It is at least arguable that the "Theory of Conversation" is a component of the "Theory of Meaning". And even if this interpretation is resisted, it is undeniable that the theories are mutually illuminating and supportive, and that they are of more philosophical, linguistic and historical interest if the temptation is resisted to discuss them in isolation from each other (Neale, 1992, pp. 511–512).

I endorse the view that the theory of implicature derivation is (or at least should be) a component of an inferential theory of communication. There are two points that I want to underline about inferential-intentional pragmatics in this section:

(1) A key to Grice's work on meaning is that the intentions behind speakers' utterances play a causal role in hearers' inferring speaker's meaning. In terms of Grice's taxonomy of reasons, this means that *personal,* or *justificatory-explanatory* reasons are the kind of reason hearers have for their interpretative responses to what speakers utter. On currently standard assumptions about actions, speakers' intentions also rationalise and cause their utterances, given the meaning they intend to convey.

(2) I argue that making sense of utterances counts as reasoning, whether it is conscious or not, and whether it involves heuristics or not. I suggest that an implicit commitment to this view on Grice's part is indicated by parallels between Grice's discussion of calculability and his view that reasoning can take place quickly and implicitly.

As I explained in chapter 2, Grice did not think that inference or reasoning is always conscious and explicit. There is a "'hard way' of making inferential moves" which is "[a] laborious, step-by-step procedure [that] consumes time and energy", and there is "A substitute for the hard way, the quick way, ... made possible by habituation and intention" (Grice, 2001, p. 17).

It is my thesis that pragmatic processing is full-blown reasoning even though it typically proceeds the 'quick way'. In this chapter, I put to use the claim made in chapter 2 that reasoning is inference undertaken in pursuit of a goal. This, I argue, is how pragmatic interpretation proceeds. Pragmatically derived material is mostly arrived at 'the quick way', where the quick way involves heuristic processes and, often, shortcuts. This is nonetheless reasoning, since it is still inference towards a goal directed by a purpose. In this case, the purpose is to work out what the speaker meant by her utterance. The pursuit of this goal is 'wired in' to the systems used for utterance interpretation.

My opinion is in contrast with views according to which reasoning or inference is necessarily conscious in some sense. If one holds this opinion and wants a Gricean story about pragmatics, then one has four options. One could claim that the processes involved in inferential pragmatics are typically conscious and explicit. The evidence, however, is that we are not generally aware of making step-by-step Gricean derivations. This leaves three alternatives, all of which recognise that pragmatic processing is at least often fast and subliminal. (1) One can take Gricean explanations as constraints on the interpretations reached, but without any pretension to be descriptions of the process by which the interpretation is derived. (2) One can say that only conscious pragmatic processes are properly inferential, and that unconscious processes are merely heuristic, routinised versions that mostly proceed *as if* they were inferential. (3) One can say an agent need not be aware of reasoning at the time it takes place, but it is characterised by its availability to consciousness. That is, reasoning can always be reconstructed after the fact, and Gricean explanations are such reconstructions. This last is François Recanati's view. I return to consideration of these alternatives in section 4.3.3, after considering the role of reasons in Grice's theory of meaning.

4.3.2 REASONS AND CAUSES

Reasons and Grice's theory of meaning

Meaning is grounded in reasons in Grice's work. For a speaker, S, to mean proposition p by addressing an utterance x to a hearer A, S has to intend that A comes to think that S believes p, and that A comes to think this at least in part *because* of S's utterance of x. What this 'because' comes down to is that S's utterance of x must provide A with reason(s) to think that S believes x. In this section I analyse this theory, drawing on Stephen Schiffer's detailed and careful account of Grice's views (in Schiffer, 1972) as well as Grice's own exposition.

As Schiffer says, there are two conditions which must be met for an utterance *x* to mean something in Grice's sense. The first is that:

> S must intend to produce [response] *r* in [the hearer] A "by means of" A's recognition of S's intention to produce *r* in A. (Schiffer, 1972, p. 10)

Here *r* is the hearer's response to the utterance. For ordinary assertions, the response aimed at is the belief that the speaker believes *p*[167], but Grice's formulation is general enough to allow for other possibilities, providing considerable flexibility in the type of meanings that imperatives and interrogatives, as well as assertions, might have. The first condition makes the claim that recognition of the speaker's intention plays a causal role in bringing the hearer to his interpretation of the utterance, as Schiffer explains.

> If we allow that reasons are causes, we may say that S intends *r* to be produced in A by virtue (at least in part) of A's belief that S uttered *x* intending to produce *r* in A just in case S uttered *x* intending that A's belief that S uttered *x* intending to produce *r* in A be (at least) a necessary part of a sufficient cause of A's response *r*. (Schiffer, 1972, p. 10)

I explore the point that this kind of explanation involves both justification and causation, that is, reasons that are also causes, below. This plays a crucial role in the second condition:

> The other restriction is that A's belief that S uttered *x* intending to produce *r* in A must not merely be intended to be a cause of A's response *r*, it must also be A's reason, or part of A's reason for A's response *r* ... (Schiffer, 1972, p. 10)

This means that arriving at speaker's meaning is a matter of arriving by reasoning – this comes from the second restriction – at the speaker's intended meaning, and the process is set in train by recognition of the intention behind the utterance. The pattern that the inference follows is set out by Schiffer:

[167] In a mental-representation theory, a mental representation of the speaker representing *p* as true.

> What Grice had in mind was simply this: sometimes the fact that a certain person believes (or believes he knows) a certain proposition to be true is good evidence that that proposition is true, and sometimes the fact that a certain person intends (or wants) another to believe that a certain proposition is true is good evidence that the former person himself believes (he knows) that that proposition is true. (Schiffer, 1972, p. 11)

So if Bertrand says to Ludwig: "The cat is on the mat", then Ludwig may recognise that Bertrand intended him to believe that he (Bertrand) believes that the cat is on the mat; and that may be good enough evidence for Ludwig to infer that Bertrand believes that the cat is on the mat. This much is certainly part of communication, in the strict sense. Ludwig *may* then go on to infer that the cat is on the mat, depending on his attitude towards Bertrand's beliefs and general epistemic state. This may or may not be part of the communication, depending on Bertrand's intentions. If Bertrand only intends to 'exhibit' his belief that the cat is on the mat (and let Ludwig draw his own conclusions about whether that state of affairs actually holds) then the utterance was 'exhibitive'. If, on the other hand, the response that Bertrand (primarily) intended to produce in Ludwig was the belief that the cat was on the mat, then the utterance was 'protreptic'.[168] Whether any particular assertion is protreptic or exhibitive, and whether assertions generally belong to one class or the other, are interesting questions which I pass over here. Either way, this picture of communication is essentially inferential: the hearer's response is derived by reasoning from the speaker's intention. This is in contrast with such theories as Millikan's non-inferential view, mentioned in chapter 1, according to which the hearer comes to believe that (e.g.) the cat is on the mat, but this belief is reached in a way that does not depend on the speaker's intentions at all, being in that respect more akin to perception than reasoning.

Grice's central point is that recognition of the intention behind an utterance is – at one and the same time – *evidence* that the speaker thinks *p*, (i.e.

[168] Grice made a distinction between exhibitive utterances "utterances by which U M-intends to impart a belief that he (U) has a certain propositional attitude" and protreptic utterances "utterances by which U M-intends, via imparting a belief that he (U) has a certain propositional attitude, to induce a corresponding belief in the hearer" (1989c, p. 123) this is from chapter 6, originally published as (Grice, 1968).

provides reason to think that the speaker thinks *p*) and is the *cause* of the hearer's coming to believe that the speaker thinks *p*. As Grice wrote in his first published paper on the subject, "... in some sense of 'reason' the recognition of the intention behind [an utterance] *x* is for the audience a reason and not merely a cause." (Grice, 1957, p. 385)

This formulation rules out certain cases where an utterance produces an involuntary response in the hearer, as Grice explained at the time. (See also Schiffer, 1972, p. 8.)

> Suppose I discovered some person so constituted that, when I told him that whenever I grunted in a special way I wanted him to blush or to incur some physical malady, thereafter whenever he recognized the grunt (and with it my intention) he did blush or incur the malady. (Grice, 1957, p. 385)

As Schiffer says, the blush would not count as the meaning of the grunt. "Should he then grunt, we should not, Grice thinks, want to say that he thereby meant something." (Schiffer, 1972, p. 8) Taking Grice's theory as the foundation of an inferential account of communication, we can say that this strange state of affairs would not be a case of communication. Communication is limited to cases where the intention behind a speaker's utterance justifies the hearer's response as well as causing it.

One might ask how Grice thought that the intention behind an utterance was to be recognised. The answer is that it can be worked out, based partly on the usual meaning of the words and expressions used, and partly on the context:

> in cases where there is doubt, say, about which of two or more things an utterer intends to convey, we tend to refer to the context (linguistic or otherwise) of the utterance and ask which of the alternatives would be relevant to other things he is saying or doing, or which intention in a particular situation would fit in with some purpose he obviously has (e.g., a man who calls for a 'pump' at a fire would not want a bicycle pump). Non-linguistic parallels are obvious: context is a criterion in settling the question of why a man who has just put a cigarette in his mouth has put his hand in

> his pocket; relevance to an obvious end is a criterion in settling why a man is running away from a bull. (Grice, 1957, p. 387)

In this passage is the germ of Grice's work on inferring implicatures of utterances, itself the starting point for modern inferential pragmatics. As discussed in chapter 1, for almost every utterance of any phrase there will indeed be two or more things (generally many more) that a speaker might have intended to convey, so 'recognition' of the intention behind an utterance must be an inferential affair, guided by context and, in most cases strongly aided by the hearer's knowledge about "what is normally conveyed" (Grice, 1957, p. 387) by the expressions used.

For Grice, then, the intention behind an utterance is inferred from the expressions used and the context, and is both a cause of and a reason for the hearer's interpretation. I suggest below that this second point means that the kind of reasons needed for Grice's theory of meaning are those he described as personal or justificatory-explanatory reasons in his later work on reasoning.

Causalism

There is a link to Donald Davidson's well-known work on intentional actions (1963; 1980a), and the related causal theory in epistemology. The key point that Davidson was trying to establish was that an agent's reasons for an action are causally effective: there is no bar on identifying something as both a reason for and the cause of an action, and indeed for intentional actions the operative reasons are distinguished by the causal role they play in the action from other reasons that there might be for that action.

Causalism about actions makes two claims. First, "An event's being an action depends on how it was caused" (Mele, 1997a, pp. 2-3), and, secondly, actions are to be explained in terms of psychological or mental events such as beliefs, desires, intentions (Mele, 1997a, pp. 2-3).

This is hardly a modern theory. According to Aristotle, "the origin of action – its efficient, not its final cause – is choice, and that of choice is desire and reasoning with a view to an end." (Aristotle, 1998, p. 139: 1139a, 31–2) Aristotle's theory is that the choice of an action is a causal explanation of the action, rather than a directly teleological explanation – an explanation in terms of what makes an action happen, what brings it about, rather than directly in terms of what purpose it serves. There is, however, a teleological aspect to this kind of causal explanation, since the causes it posits for actions are choices which derive from reasoning about goals. That is, given a desired outcome, choice of action comes from reasoning about the kind of action that is likely to achieve it. Aristotle's was an account of action in terms of causes which are founded in an agent's reasons for action.

Causalism has become the "standard theory of action" in recent decades (Schlosser, 2007, p. 187), largely on the basis of Davidson's argument that there has to be some way of distinguishing between reasons that an agent has for an action (possibly without being aware of them) but does not act on, and the reasons that are actually operative[169]. Assuming that both sides of the debate accept that when agents act intentionally they act for reasons, Davidson's challenge to non-causalists was to provide "an account of the reasons *for which* we act that does not treat those reasons as figuring in the causation of the relevant behaviour" (Mele, 1997a, pp. 11, his italics)

Causes for beliefs

The related causalist theory about beliefs has also become standard in philosophical accounts of belief-formation and of reasoning. The account of reasoning as value-preserving transitions developed in chapter 2 is of this type: in reasoning the premise mental states give rise to subsequent conclusion mental states. Ralph Wedgwood summarizes causalist views of epistemology in a recent paper:

> If your reason for forming a certain belief is 'represented' by some of your antecedent mental states, then your formation of that belief is – as epistemologists often put it – 'based on' those antecedent mental states. Like most contemporary epistemologists, I take this 'basing relation' to be a kind of causal relation: for your formation of this new belief to be based

[169] Also on the basis of Davidson's devastating replies to standard objections to causalism, among them (1) the idea that such explanations are flawed because causes are not logically distinct from the actions they cause and (2) that one cannot provide causes of action that are both necessary and sufficient. Schueler, himself a non-causalist, thinks that these arguments were demolished by Davidson (Schueler, 2001, pp. 263, fn 3).

> on those antecedent mental states, you must have formed that new belief precisely because you were in those antecedent mental states – where this is the 'because' of ordinary causal explanation. (Wedgwood, 2006, p. 661)

The Gricean account of the hearer's end of the conveying of speaker meaning fits squarely with what has come to be the modern orthodoxy: the hearer's recognition of the speaker's intention causes the response in the hearer that is the speaker's meaning. On the speaker's side, we may say that for the speaker, making the utterance she did, in the way that she did, her intention to evoke a certain response (which we take as given, as discussed in section 4.2 above) was both a reason for and a cause of her action.

Two frequently raised problems for causalist accounts need not worry us. The first is the much-discussed question of deviant causal chains. It is important for causalist theories to be able to distinguish a reason that is a cause (for a belief, an action or an intention) *in the right way*, from other reasons for action that are causally effective in other ways. A belief *p* that is a reason for forming belief *q* might causally lead to its formation, but in a way that has nothing to do with justifying it. Holmes might have come to believe that a snake was the cause of death because his belief that the hole, the whistling sound and the bell-pull were significant led him to write a despairing letter to his brother Mycroft, who wrote back with the solution to the case. Here Holmes' belief that the snake was the means of death is caused by a set of prior beliefs that constitute evidence, but not in the right way for that belief to be his reason for that conclusion (cf Wedgwood, 2006, p. 663).

We already have a solution for this kind of problem. The 'right way' for a cause to rationalise a belief is via a reasoning process with the reason as the input and the belief as the output, so that they are related as premise and conclusion in an inference (demonstrative or non-demonstrative). (Wedgwood, 2006 is a detailed attempt to show that an account of reasoning of this kind deals with the problem of deviant causal chains for beliefs.) A belief *p* which causes belief *q* via such a process is an *operative* reason for belief *q*. This criterion, as I have noted in chapter 2, rules out associative connections. Suppose Holmes' belief that the whistle was significant reminded him of a pub called *The Pig and Whistle*, and thinking about pigs reminded him of a different piece of evidence that he had forgotten and that was a reason to believe that the cause of death was a snake, and he came to believe so on that basis (cf Wedgwood, 2006, p. 667). The connection between the original belief about the whistle and the conclusion reached is not by reasoning, and so our account of reasoning allows us to rule that while the original belief is part of the causal chain leading to the correct conclusion, and a (potential) reason for reaching it, it was not an operative reason.

If, as Grice suggested, practical reasoning is also primarily a matter of value-preserving transitions, then a similar story can be told for the way that a speaker's intended meaning leads to her utterance. Whatever one might think of the prospect of this kind of explanation holding for practical reasoning, this is not a problem specific to a theory of communicative action, and it is perhaps too much to expect it to be solved within pragmatics.

A second question for causalist accounts which has attracted a fair amount of recent debate is whether a reason is a mind-independent fact or a mental state. Wedgwood's solution, which I happily adopt, is to use the fruitfully ambiguous formulation that a premise mental state *represents* the reason for a conclusion mental state. The ambiguity cannot matter. Whether, strictly speaking, the *fact* that there is water on the ground is the reason for an agent's belief that it rained last night, or it is the agent's *belief* that there is water on the ground that plays that role, one must still have had the premise belief to have reached the conclusion.[170]

According to causalism about actions, any intentional action has a cause that is a reason for that action. Similarly, causalism about beliefs claims that a belief reached by reasoning will have a cause that also is a reason for that belief. Davidson called an explanation of action in terms of the agent's reason for doing what he did a rationalisation. Davidson argues that giving an agent's primary reason for an intentional action is a way of explaining the action causally: "rationalization is a species of causal explanation" (Davidson, 1963, p. 3).[171] Every rationalization justifies, in what Davidson calls an "irreducible –

[170] I have changed the example from Wedgwood's – frost as a reason to believe it was freezing last night – since frost is evidence that it is freezing now.

[171] This use of the word is in contrast to its dominant normal use, where a rationalization is an explanation that a person concocts for his action after the fact, giving a reason because it would be convenient if that reason had caused the action, rather than because that reason was actually operative. Spurious explanations given by participants in hypnosis for actions

though somewhat anaemic – sense" (Davidson, 1963, p. 9). That is, "from the agent's point of view there was, when he acted, something to be said for the action." (Davidson, 1963, p. 9) – generally that it was believed to be a means to the realisation of some goal towards which the agent had a 'pro-attitude' (a desire, yen or similar). Rationalisations also explain. Rationalisation is a type of causal explanation, "distinguished from other causal explanation by possessing the property of justification" (Davidson, 1963, p. 9).

According to a Gricean theory, rationalisations of this sort are a key aspect of pragmatics, since, as discussed above, the hearer's recognition of the intention[172] behind an utterance is both the cause of and a reason for the hearer's response.

Justificatory-explanatory reasons

The notion of a cause which is also a reason is also one of the clearest links between Grice's work on meaning and his work on reasoning, where he distinguishes three different types of reason: *pure explanatory*, *justificatory* and a third, hybrid, type, *justificatory-explanatory*. (Grice, 2001, ch.s 2 & 3) It is the third type, the justificatory-explanatory or *personal* use, I think, that is the kind of reason Grice works with in his theory of meaning.

Type 3 reasons can be expressed in sentences of the form "X's reason(s) for A-ing was that B (to B)". (Grice, 2001, p. 40) For example, "John's reason for thinking Samantha to be a witch was that he had suddenly turned into a frog." (Grice, 2001, p. 40)

Type 3 reasons are simultaneously explanatory and justificatory: "they explain, but *what* they explain are actions and certain psychological attitudes" (Grice, 2001, p. 41, his italics). They are justificatory, in the sense that B *seems to* X to justify A (B may or may not *actually* justify A) (Grice, 2001, p. 41). That is, they are justificatory precisely in Davidson's anaemic sense.

Grice discusses whether type 3 reasons are causes "of that for which they are reasons" (Grice, 2001, p. 44). He points out that the debate is about causal explanation, and suggests that an objection to that theory based on ordinary

carried out under post-hypnotic suggestion are examples.

[172] Note that Grice's phrase is usefully non-committal on whether it is a mental state or a mental event which plays the role of rationalising cause.

usage of the word 'cause' is beside the point (Grice, 2001, p. 41). In ordinary use type 3 reasons are not causes, he claims. For example, "My love of cricket may cause me to neglect my work, but did not (in the vernacular sense of "cause") cause me to play yesterday." (Grice, 2001, p. 44) Grice is of course well-known for arguing that usage does not map simply onto word meaning, although he did not do so explicitly in this instance. It may be that one would not say "My love of cricket caused me to play yesterday" because it is odd to say so in those terms, rather than because it is false.

What is more important is whether a correct explanation of the action would include the love of cricket (a pro-attitude, in Davidson's terms) as a cause:

> the debate is not really about whether reasons are causes in the vernacular sense; it is about whether to specify a type (3) reason as the explanation of an action is to give a "causal explanation" of the action, in a sense of "causal explanation" which is none too clear to me, and which (I sometimes suspect) is none too clear to the disputants. (Grice, 2001, p. 44)

There is a further point of congruence between what Grice says about type 3 reasons and what he says about reasoning as it relates to meaning. Comprehension of speaker meaning is often accomplished without conscious, explicit reasoning. So if type 3 reasons are, as I claim, the kind of reasons that hearers have for the meanings they derive from utterances, they must be capable of acting as personal reasons unreflectively. That is, it must be possible to come to a particular understanding of an utterance owing to a type 3 reason, that is, with one's interpretation justified somehow by the intention behind an utterance and caused by it, but without necessarily being explicitly, consciously aware that the intention justifies the interpretation. Grice's discussion of type 3 reasons mentions just this kind of possibility[173]:

> ... if X's reason for A-ing is that B ... it is necessarily the case that X regarded (*even if only momentarily or subliminally*) the fact that B in justifying him as A-ing. (Grice, 2001, p. 41, my emphasis.)

[173] Although Grice does not relate this point to his work on utterance interpretation.

4.3.3 GRICEAN INFERENCE AND EXPLANATORY POWER

Calculability of conversational implicatures

We have seen that Grice was committed to understanding use of language as a rational activity, in which a hearer's interpretation of an utterance was rationally grounded in the intention behind an utterance, and that intention could be worked out from what was uttered and the circumstances of the utterance. The recognition of the intention behind production of an utterance – an intention, for example, that he come to think that the speaker believes p (in exhibitive cases) and that he come to think this (at least partly) as a consequence of the speaker's making the utterance – provides the hearer with a reason to form that belief.

The aim of seeing language use as grounded in reasoning is particularly clear in Grice's insistence on the calculability of conversational implicatures. I therefore turn now to this second aspect of Grice's inferentialism as it relates to language use, the role of inference in his theory of conversation, specifically the derivation of implicatures, bearing in mind Neale's point that Grice's theory of conversation can be seen as a component of his theory of meaning.

As is well-known, in his work on conversation Grice showed that the meaning that a speaker conveys by making an utterance on some occasion may go well beyond what is asserted, or what is expressed in virtue of the 'timeless' normal meanings of the expressions used[174]. Utterances can have implicatures – implications which are part of the intended meaning of the utterance – as well as explicit content. Grice proposed that conversational implicatures can be worked out from what is said (and the way it is said) by assuming that the speaker is conforming to the Cooperative Principle and (at least some of the) conversational maxims.

> The presence of a conversational implicature must be capable of being worked out; for even if it can in fact be intuitively grasped, unless the intuition is replaceable by an argument, the implicature (if present at all)

[174] These are, approximately, the two different senses of Grice's 'what is said'. For discussion of Grice's settled view of 'what is said' see Wharton, 2002.

> will not count as a CONVERSATIONAL implicature; it will be a CONVENTIONAL implicature. (1975, p. 50, his emphases.)

> ... the final test for the presence of a conversational implicature [has] to be, as far as I [can] see, a derivation of it. One has to produce an account of how it has arisen and why it is there. (1981, P. 187)

That is, there are 'conventional' aspects of meaning (implicit as well as explicit) whose recovery is a matter of knowing and retrieving the relevant meaning, but crucially, non-conventional components of the meaning of an utterance must be inferred, worked out rationally. Whenever there are non-conventional components, there must be a derivation of such elements in the form of an argument. Again, it is crucial, as with Grice's work on speaker meaning, that this derivation explains for each implicature not only the reason "why it is there", but also provides some kind of account of "how it has arisen" – that is, a causal account.

Grice's schema

Grice's schema for the derivation of conversational implicatures shows that the process is envisaged as inference to the best explanation, where what is to be explained is that the speaker has said that *p* (in a certain way, in particular circumstances), and the explanation sought is in terms of the speaker's intention to convey an implicature *q*:

> A general pattern for the working out of a conversational implicature might be given as follows: 'He has said that *p*; there is no reason to suppose that he is not observing the maxims, or at least the CP; he could not be doing this unless he thought that *q*; he knows (and knows that I know that he knows) that I can see that the supposition that he thinks that *q* IS required; he has done nothing to stop me thinking that *q*; he intends me to think, or is at least willing to allow me to think, that *q*; and so he has implicated that *q*."' (Grice, 1975, p. 50: his emphasis.)

The Relevance Theory schema

Related patterns can be given in other inferential pragmatic theories. For example, Sperber and Wilson give the outline in table 4 (below) for Peter's inference of Mary's intended meaning when she makes the utterance in (20b). (Sperber & Wilson, 2004, pp. 615–616: their examples (11) and (12).)

(20) a. Peter: Did John pay back the money he owed you?

b. Mary: No. He forgot to go to the bank.

Here the inference is guided by the presumption of optimal relevance and situation-specific expectations of relevance rather than the CP and maxims, but there are similarities between the two schemas. In both the Gricean and the relevance-theoretic schemas, the input is something that the speaker has uttered, and the manner and circumstances of its utterance; the output is a hypothesised component or components of the speaker's meaning that serves to explain why the utterance was made in the way that it was. The output meaning is, in both cases, in the form of an intention attributed to the speaker. In both schemas, the output is inferred on the basis of the input, together with such extra assumptions as are necessary, given a standing presumption or presumptions about the standards that the speaker's utterance will attain.

There is, however, a significant difference here. In Grice's schema, observance of the CP and maxims supplies premises to the argument[175]; in relevance-theoretic derivations the direction of the whole derivation – and when it stops – are guided by expectations of relevance, and governed by a presumption that the utterance will be optimally relevant. I comment in chapter 5 on these points and the comprehension procedure they mandate.

A further difference is that Sperber and Wilson's example illustrates three interlocked inferences to the best explanation. Peter infers (1) Mary's explicit meaning, (2) an implicated premise and (3) her implicated conclusion and perhaps other weak implicatures. In Grice's schema only an implicature is derived. However, applying Gricean reasoning to an example such as (20), one

[175] Sperber and Wilson discuss this point (Sperber & Wilson, 1986, p. 36). As they say, others have recast Grice's maxims as code-like rules.

Table 4: Example outline of relevance-theoretic comprehension

(a) Mary has said to Peter, "He$_x$ forgot to go to the BANK$_1$/BANK$_2$." [He$_x$ = uninterpreted pronoun] [BANK$_1$ = financial institution] [BANK$_2$ = river bank]	*Embedding of the decoded (incomplete) logical form of Mary's utterance into a description of Mary's ostensive behaviour.*
(b) Mary's utterance will be optimally relevant to Peter.	*Expectation raised by recognition of Mary's ostensive behaviour and acceptance of the presumption of relevance it conveys.*
(c) Mary's utterance will achieve relevance by explaining why John has not repaid the money he owed her.	*Expectation raised by (b), together with the fact that such an explanation would be most relevant to Peter at this point.*
(d) Forgetting to go to the BANK$_1$ may make one unable to repay the money one owes.	*First assumption to occur to Peter which, together with other appropriate premises, might satisfy expectation (c). Accepted as an implicit premise of Mary's utterance.*
(e) John forgot to go to the BANK$_1$.	*First enrichment of the logical form of Mary's utterance to occur to Peter which might combine with (d) to lead to the satisfaction of (c). Accepted as an explicature of Mary's utterance.*
(f) John was unable to repay Mary the money he owes because he forgot to go to the BANK$_1$.	*Inferred from (d) and (e), satisfying (c) and accepted as an implicit conclusion of Mary's utterance.*
(g) John may repay Mary the money he owes when he next goes to the BANK$_1$.	*From (f) plus background knowledge. One of several possible weak implicatures of Mary's utterance which, together with (f), satisfy expectation (b).*

could use the CP and maxims to disambiguate (Sperber & Wilson, 1986, p. 34) and perform reference assignment, operations that contribute to the explicit content of the utterance.

Indeed we have seen that Grice thought that in cases of ambiguity at least, inferences based on the context of utterance would be needed to determine which possible meaning was intended. In fact, the principles regulating pragmatic processing (whatever they are) must also be active in the inferential work necessitated by ellipsis, vagueness, loose use and other ways – beyond ambiguity and referential indeterminacy – in which linguistic meaning falls short of the proposition expressed by an utterance (Wilson & Sperber, 1981; Neale, 1992, pp. 520, n 27).

4.3.4 EXPLANATORY POWER

Explanation and justification

As I discussed above, the question is often raised how schemas of this type can be explanatory, given that hearers are not, most of the time, aware of constructing or rehearsing arguments of this sort in communication. Instead, the typical experience is that the speaker's meaning (or rather, the hearer's best estimate of it) is immediately available to the hearer without any need for conscious, explicit reasoning. It has been seen as problematic for Gricean explanations that explicit inference is largely absent from our experience of utterance interpretation. What, then, is the relation of such schemas to what goes on in the mind of a hearer?

A subsidiary question is what Grice's own view was. Taking the question about Grice first, one interpretation, which is fairly clearly mistaken, is that Grice thought that participants in conversation have to consciously, laboriously work their way through the derivation of what is meant from (facts about) the utterance and some principles of rational cooperation, passing through the mental states in the derivation, with awareness of doing so. It is straightforward to see that if this were so, schemas of this type could answer both the *how* and the *why* questions: the process would amount to building an argument step by step (including some non-demonstrative steps, presumably), as in Grice's picture of reasoning. The input would thus both rationally justify and bring about the output. However, it is clear from what Grice says that this was not his view: implicatures can be "intuitively grasped" (Grice, 1975, p. 50).

A more plausible – and, I think, widespread – interpretation of Grice is that he thought that sometimes reasoning or inference is involved in arriving at implicatures and sometimes it is not. When it is not, the implicature is grasped 'in a flash', intuitively. In these cases, one can always construct a chain of inferences which show how reasoning *might have* proceeded if there had been any reasoning involved, as is required by calculability, but, according to this view, in fact, on these occasions, there was none. Richard Warner interprets Grice in this way, as I discuss below.

In this section, I argue instead that schemas of this kind are explanatory in part because in understanding utterances, hearers are engaged in reasoning, although they are not typically aware of the process. I also suggest that there are indications that Grice may have held this view, so that when a conversational participant arrives at an implicature, the process that got him or her to the implicature would count as reasoning for Grice. That is, I do *not* want to argue that Grice thought that on all occasions when language was in use speakers and hearers had to be engaged in *explicit, conscious* reasoning (that is the first of the three views), but I do want to suggest that he thought that they were engaged in reasoning.

In the end, regardless of Grice's position, this is the view that I take. A view of this sort links together the answers to the *how* and the *why* questions. The schema shows why the interpretation is reached in that it shows that the interpretation was justified, in Davidson's 'anaemic sense', at least, that it shows that there were reasons that seemed to the hearer at the time to justify the interpretation reached. The derivation, in other words, shows that there were reasons for the hearer as a rational but fallible agent to reach a certain interpretation.

The schema also shows in a certain sense how the interpretation was reached. It presents a reasoning process and the claim is that it was *that* reasoning process that led from the utterance to the interpretation. There is a problem with this claim, however. It does not fully answer the question, "In what way are derivations according to such schemas explanatory?"

There are two ways of making this objection, one of which is more cogent than the other. The less cogent way is as follows: We know (the objection claims) that such schemas are not causally explanatory because we are not aware of going through the steps of the derivation. Then the question about the explanatory force of Gricean derivations becomes the following:

> what is the relation between the reasoning you *might* have engaged in and your understanding the sentence? How is there any explanatory power in the fact that, although you reached your understanding of the sentence *in some other way*, you *might* have reasoned your way to such an understanding? (Warner, 2001, p. x, his italics)

This objection is itself vulnerable to an objection. We know that unconscious processing goes on all the time, including, if mental-logic theorists are right, series of transitions between conceptual representations which parallel arguments. Why then should broadly Gricean derivations not be instantiated in processing literally, step by step, but unconsciously? So this version of the objection would fail unless it could be shown that there could not be unconscious derivations of this sort, and I do not think that this case has been made.

As we shall see, François Recanati would have a different objection to this first line of argument. For him, all inferential pragmatic processes are conscious in what he regards as the important sense: that they are available (at least retrospectively) to reflective awareness. I discuss Recanati's position further, below.

A better version of the objection, in my view, runs as follows: There are good reasons to suppose that heuristic processes are used in thought, particularly rapid thought, conscious or unconscious, so it is implausible to believe that thought processes isomorphic to Gricean derivations take place in all or even most cases of utterance interpretation. If Gricean[176] inferential schemas do not describe the thought processes and mental states involved in arriving

[176] Note that while the relevance-theoretic type of derivation seen in table 4 is a good deal closer to the underlying heuristic it may also be somewhat idealised: no interpretations or parts of interpretations which were generated and rejected are mentioned. Further discussion of the relevance-theoretic comprehension procedure is reserved for chapter 5.

at an implicature, then in what sense, this version of the objection asks, could they provide an explanation of how the implicature is derived?

I take it that Millikan's objection to inferential, Gricean pragmatics discussed in chapter 1 is along related lines. Her objection is that the "Gricean analysis is very implausible if taken at face value as requiring that speakers and hearers harbor multiply embedded mental representations of one another's mental representations during normal conversation". (Millikan, 2005, p. 203) I assume that the objection here is that the representations postulated are improbable because they are too complex, rather than because we are not consciously aware of tokening such representations in utterance interpretation – and thus this is related to the second objection.

I do not agree with Millikan's objection, however. The complexity of representations in itself need not be any bar to processing. We have considerable facility with metarepresentations of thoughts and utterances, up to three or four levels deep. (Consider how many conversations include assertions along the lines of "He thought that she said that he said that...".) What is more, a good deal of the embedding can be done automatically: the parsed structure of an utterance is presumably automatically embedded under: S *[the speaker] said* "...". Similarly, the output of utterance interpretation procedures, even when they are very simple heuristics, as with 'naive optimists', is presumably embedded under: S *means that*.... So I do not think that the need for embedded representations would tell against a full-computation version of Gricean accounts. Rather, I think it is implausible that full derivations take place, with manipulations of mental representations according solely to value-preserving rules. I think that it is important to answer this form of the objection to Gricean derivations.

I have already laid the ground for my answer to this version of the objection in chapters two and three. My answer, which I also attribute to Grice, is that most utterance interpretation proceeds a quick, heuristic way and is related to an explicit step-by-step derivation in just the same way that reasoning the quick way is related to reasoning the hard way. In fact, since I claim that working out the interpretation of an utterance is a reasoning process, I am claiming that the former cases are a subset of the latter.

I am claiming also that Grice was implicitly committed to the position that language use involved reasoning, regardless of whether the thought processes involved on a particular occasion were conscious or not. This seems to me to emerge from comparing what Grice said about aspects of language use with his views about reasoning in general. Therefore I discuss here the parallels between what Grice said about the quick way of reasoning and about the need for implicature derivations to be reconstructable.

We have seen that Grice did not think that arriving at implicatures always involved conscious explicit inferences. Sometimes one might work out an implicature laboriously; sometimes one might grasp it intuitively. Similarly, Grice did not think that reasoning in general was always conscious and explicit. As discussed at length in chapter 2, his opinion was that there is a hard way of reasoning, which is the laborious, step-by-step construction of inferential chains, and there is a quick way, which is a substitute for the hard way, and is possible because of the intention behind the reasoning as well as habituation to particular kinds of inferential move (Grice, 2001, p. 17). What is important is that the 'quick way' of making inferential moves also counts as reasoning. Grice is quite clear about this. For example, he says that in the absence of explicit reasoning,

> The possibility of making a good inferential step (there being one to be made), together with such items as a particular inferer's reputation for inferential ability, may determine whether on a particular occasion we suppose a particular transition to be inferential (and so to be a case of reasoning). (Grice, 2001, p. 17)

The parallel with what Grice says about calculability is exact. A mental or verbal transition intuitively made will count as a case of reasoning if it parallels an inferential step. If it does parallel an inferential step, then in principle it is capable of being worked out, just as "the presence of a conversational implicature must be capable of being worked out... even if it can in fact be intuitively grasped" (Grice, 1975, p. 50). The conclusion that I draw from this is that the parallel is so exact that Grice was implicitly committed to the view that I am advocating: arriving intuitively at a conversational implicature is an instance of reasoning.

This recasts Warner's question about the explanatory status of Gricean derivations as follows: *what is the relation between the fully explicit reasoning that you* might *have engaged in and understanding the sentence? How is there any explanatory power in the fact that, although you reached your understanding of the sentence by reasoning* the quick way, *you* might *have reasoned* the hard way *to such an understanding?*

The claim that the explicit derivation is a causal explanation in such cases of quick, intuitive interpretation rests on the claim about quick reasoning in general that it is reasoning in that it is aimed at resembling reasoning the hard way. An intention, or the aim/purpose of a mental sub-system, directs non-algorithmic, possibly non-value-preserving processes towards a value-preserving answer. Of course, such merely heuristic procedures do not guarantee the production of the answer that an explicit reasoning process, parallel to an argument, would have reached. To understand the form that pragmatic processing takes we need to investigate the details of the heuristics used, as well as the explicit inferential derivation whose result the heuristics are aimed at reproducing. I suggest some answers to these questions in some remarks in the remainder of this section and in chapter 5.

The view that I am taking is quite closely related to the position taken by Sperber and Wilson. For them, pragmatic processing is inferential, whether it is spontaneous or laborious. Sperber (1995) suggests that the term 'inference' is used by psychologists because it avoids the connotations of conscious explicitness that the word 'reasoning' has. There are subliminal, spontaneous inferences as well as conscious ones (Sperber, 1997); but the more important distinction for psychological explanation is that between inferential and non-inferential (including pseudo-inferential[177]) processes. For Sperber and Wilson, an inference must have input and output related as premises and conclusion are in an argument (Sperber & Wilson, 1987b, p. 737). An interpretation of an utterance constructed by the hearer is an inference about the meaning the speaker intended to convey, based on (and logically supported by) the utterance. I agree with these points about inference, while adding that it is my opinion that inferential processing which is directed towards a particular goal is reasoning. It is relevant, I think, that it is not customary in the psychology

[177] I discuss pseudo-inference in section 4.3.5 below.

of reasoning to limit investigation to conscious processes, nor to say that a process used in solving reasoning problems is merely inference if it is not conscious.

Among those who favour a broadly Gricean approach, there have been several other answers to the question of how such schemas can be explanatory. One line that has been taken is that the Gricean schemas have no psychological reality as an account of the processes involved in utterance interpretation. According to this view, they simply express constraints on correct interpretations of utterances. Hearers are disposed somehow to reach interpretations which satisfy the CP and maxims, or the Presumption of Relevance, but Gricean inference schemas do not describe how such interpretations are reached. (One recent advocate of such a view is Båve (2008).) A related view is that Gricean derivations have normative force rather than psychologically descriptive force (Saul, 2002a; see also Saul, 2002b for Saul's views on psychological reality and Gricean explanations)[178]. Such views suggest that Gricean schemas are not causally explanatory: they explain *why* (in one sense or another) but not *how*. Those who hold this kind of view would say that the question of how interpretations are actually reached is a separate psychological issue.

Psychological reality and consciousness

Other theorists have focussed on the question, less important, in my opinion, of the apparent mismatch between Gricean derivations and our introspective intuitions about pragmatic processing. I think two broad lines can be distinguished here. One line is to deny that pragmatic processing generally involves inference or reasoning. Then Gricean derivations explain only in the sense that they illustrate how reasoning would proceed if there were any: they are 'as if' explanations. Recanati's view is the converse. For Recanati, implicature derivation is conscious: since it consists in person-level inferences it could not, he claims, be otherwise. I have explained that I share with Sperber and Wilson the view that the crucial elements of psychologically real, broadly Gricean explanation are 1) that it involves inference about speaker's intentions, and 2) that an account is given of the processes that perform the infer-

[178] Saul holds this view herself, and attributes it to Grice, if I have understood her correctly.

ence. Whether the inferences involved are conscious seems to me to be a secondary question, at best. However, the question about the explanatory force of Gricean pragmatics has been thought to concern the availability to consciousness of explicit derivations, and I look at this question and its relation (or otherwise) to the personal/sub-personal distinction here.

One view of this question is akin to the second interpretation of Grice, above. On this type of account, some inferential pragmatics is conscious and explicit, but the conscious inference, through habituation, can become routinized and taken over by heuristics, and it is then no longer inferential. Robin Campbell (1981), for example, suggested that we should distinguish between conscious (phenic) and unconscious or subconscious (cryptic) processes. In his view, pragmatic processes are often phenic and inferential, in contrast to the exercise of linguistic knowledge, which is non-inferential and cryptic.

Campbell cites the construction of bridging inferences[179] as the kind of pragmatic process that requires conscious inference:

> Suppose, reporting a late-night gathering, someone says "And then the police arrived and we all swallowed our cigarettes". To make sense of what was said we need a bridging inference. For example, that the cigarettes contained an illegal substance. I think it is fairly clear that in general such inferences involve real cognitive effort and hence phenic structures and processes... Ordinary communication ... is littered with all sorts of repair sequences showing, or so it seems to me, effortful cognition at work. (Campbell, 1981, p. 96)

This may be so. The interpretation of novel metaphors and the comprehension of figurative devices in literature are also areas in which effortful conscious reasoning may often occur.

I would resist in Campbell's account the assumption that cognitive effort indicates conscious processing – presumably unconscious processing also requires effort – and, more generally in accounts of this type, that whether a process is conscious or unconscious tells us what kind of process it is: in particular, whether it can be inferential and whether it counts as reasoning. We

[179] The term is from Herb Clark (1977). For discussion of inferences of this type, see Wilson & Matsui, 1998; Matsui, 2000.

have seen that it appears to be Richard Warner's view that utterance interpretation is not reasoning in the absence of conscious explicit derivations. Discussing the connections between Grice's work on reasoning and his work on meaning, he writes:

> We can imagine you – the reader – reasoning as follows with regard to [a sentence, *s*]. "The sentence's standard meaning in English is [*p*]; Warner would not be producing that sentence in this context unless he intended to me to think that he believes [that *p*]. He has no reason to deceive me, so he must believe that." The problem, of course, is that people hardly ever reason this way when communicating. You did not reason in any such way when you read the sentence [*s*]. You read the sentence and understood – straightaway, without any intervening reasoning, without, indeed, thinking about it at all. (2001, p. x)[180]

Campbell, who refers to conscious pragmatic processes as 'macropragmatic' and unconscious ones as 'micropragmatic'[181], suggests that only the former involve Gricean inference:

> Macropragmatic processes would be analysed in terms of explicit inferences guided by principles of rational cooperation while micropragmatic processes would be analysed *as if* they involved such inferences. ... it may be possible to go a little further and indeed, it is desirable to do so if one dislikes the notion of unconscious inference – as I do. ... it is typically the case that these cryptic processes are merely heuristic; they deal adequately with the majority of circumstances but when they break down the control of the performance is returned by default, to deliberate phenic guidance. (Campbell, 1981, p. 100)

While I do not share Campbell's dislike of the notion of unconscious inference, there is a good deal here that I agree with. My view is that pragmatic processing is typically subliminal and that it is carried out in accordance with

[180] Warner gives a specific example, but the point he is making does not depend on it, so I have replaced his example sentence with '*s*' and the proposition it expresses with '*p*'.

[181] In my opinion a better use for this terminology would be to make the distinction between the heuristic trial-and-error search process and heuristics that are sometimes employed in the course of that search.

a heuristic, which itself typically employs further heuristic shortcuts or constraints. It is clear that when the usual process fails to deliver a satisfactory result it is possible for other, more laborious processes to step in. In cases in which we become aware of laborious processing, the additional processing is, of course, conscious in some sense.[182]

In saying this I am agreeing with Robyn Carston, who suggests that the normal state of affairs is that pragmatic processes are unconscious. However, Carston also suggests an alignment in pragmatic processing between the conscious/unconscious distinction, the distinction between modular and non-modular processes and the distinction between personal and sub-personal processes:

> The appropriate distinction within modes of processing and levels of explanation would seem to be between, on the one hand, a modular (sub-personal) pragmatic processor which, when all goes well quickly and automatically delivers speaker meaning (explicatures and implicatures), and, on the other hand, processes of a conscious reflective (personal-level) sort which occur only when the results of the former system are found wanting. (Carston, 2002a, p. 146)

I reserve comment on the modularity of pragmatic processing to chapter 5. On the application of the distinction between personal and sub-personal processes to psychological accounts of processing I think some caution is advisable, partly because I think that applying this distinction can be seen as settling by definitional *fiat* the question of whether unconscious processes can be inferential or instances of reasoning – although Carston does not employ it in this way[183]. Recanati does make this move. In his view, reasoning is a personal-level activity and therefore *must* be conscious in some sense, as I discuss below.

[182] There may be other cases in which additional processing is not conscious, such as slow unconscious processing of utterances which made no sense when first encountered. Much later the correct interpretation may spring to mind. There may have been no extra conscious processing; that does not indicate that there has been no extra processing.

[183] Carston endorses the relevance-theoretic view that there are spontaneous inferences.

The personal/sub-personal distinction

A related worry about the distinction is that it does not sit comfortably with the kind of explanation given in the cognitive sciences. Explanations are sought in terms of processes, rules and knowledge bases, perhaps modular. Such explanations will always be sub-personal in character, if it makes sense to classify them in this way at all, just as biological explanations are, whether they are in terms of organs, the properties of certain cells, or metabolic pathways[184]. Some of these explanations in terms of component systems of an organism cohere with personal-level propositions. At the personal level I might say, 'I am getting a cold, but I am fighting it'. This is presumably coherent with a sub-personal story in terms of the activity of white blood cells and various other systems and sub-systems, but neither description, in my opinion, is a substitute for the other. Similar considerations apply in psychology. That I (can) speak English (or Chinese) is a personal-level fact related to (but perhaps not reducible to or a substitute for) a scientific explanation in terms of the state of my language faculty. Here I am echoing comments made by Chomsky:

> No one expects ordinary talk about things happening in the "physical world" to have any particular relation to naturalistic theories; the terms belong to different universes... The same, then, should be true of such statements as 'John speaks Chinese'... The theory of evolution and other parts of biology do try to understand John Smith and his place in nature; not, however, under the description "human being" or "person" as construed in ordinary language and thought. (1995, pp. 32–33).

I think that the same sort of considerations apply equally to reasoning. Psychological theories of reasoning are in terms of a component system or systems of humans, and this goes as much for conscious as for unconscious reasoning. It is a fact about me as a person that I 'see' (comprehend) what you mean by your utterance, when I do. This presumably coheres with an account of the working out in terms of processes governed by rules within a system or systems, but neither account replaces the other. The descriptions serve differ-

[184] Statistical accounts, as in population biology, are quite different, of course.

ent purposes and exist on different levels. Scientific psychological explanation is conducted more or less exclusively at the level of component systems whether the phenomena to be explained are more 'sub-personal' (e.g. the workings of the visual system (Marr, 1982)) or more 'personal' (e.g. what the participant is paying attention to (Lavie & Tsal, 1994; Styles, 1997; Pashler, 1998; Lavie, 2000)). I return to this point in the discussion of Recanati's views, below.

Dual-process theories of reasoning

The use of the term 'reasoning' in psychology may differ somewhat from ordinary use, if ordinary use reserves the word for conscious, effortful activity (I am not sure that it does, but am prepared to concede the point). It has been fruitful (e.g. in mental-logic theory) to propose one system for reasoning and investigate on that basis. Thus, Rips writes that is is necessary to postulate complex unconscious processes in theories of reasoning as elsewhere in psychology:

> Johnson-Laird raises the issue of whether nonconscious procedures can be as sophisticated as conscious ones, but it is hard to see how cognitive psychology could make much progress if it were to limit nonconscious information-processing to simple routines. Surely, motor control, perception, sentence recognition, sentence production, categorization, recognition memory, and many other cognitive abilities depend on nonconscious processes of formidable complexity, and it would be astonishing if reasoning were an exception to this trend. (Rips, 1997, p. 413)

Recently, however, in the psychology of reasoning there has been considerable work on dual-process or dual-system theories. Such theories (Evans & Over, 1996; Sloman, 1996; Stanovich, 1999) posit that there are two systems for reasoning and inference.[185] System 1 is evolutionarily prior to system 2 and shared with other animals. System 2 is evolutionarily recent; presumably unique to humans. (These names for the systems are from Stanovich, 1999; Stanovich & West, 2000.) System 2 is responsible for canonical logical inference, while sys-

[185] Related claims have been made in the literatures on learning e.g. (Reber, 1993; Dienes & Perner, 1999) and judgment (Kahneman & Frederick, 2002).

tem 1 makes 'quick and dirty' approximations, by statistical processes or frugal heuristics.

For some, perhaps most, dual-process theorists (e.g. Evans & Over, 1996; Johnson-Laird, 2004)[186], the two systems are responsible for implicit and explicit inference respectively.[187] Then the claim, similar to Campbell's claim about pragmatic processing, is that normatively inferential processes are laborious and conscious, and that automatic, unconscious processes are merely heuristic: "Dual-process theorists generally agree that System 1 processes are rapid, parallel and automatic in nature: only their final product is posted in consciousness." (Evans, 2003, p. 454). Such dual-process theorists do not, of course, share Campbell's dislike of talk of unconscious inference. On the contrary, they need to talk that way in order to state the claim that unconscious inference and unconscious reasoning are carried out by non-normative processes. What dual-process theories and the Campbell/Warner view have in common is the claim that unconscious processes do not involve normative (value-preserving) reasoning: unconscious processing is merely heuristic. Since I want to argue against views of this sort in pragmatics, I make some brief remarks here about what I take to be the problems with dual-process theories of reasoning in general.

One major motivation for the system 1/system 2 distinction has been a desire to explain individual differences in reasoning ability (e.g. Stanovich & West, 1998; Stanovich & West, 2000): why, for example, do some people nearly always give the normative answer on the abstract selection task while most never do? According to a dual-process account, good performance depends on the ability to engage and use system 2, that is, the ability to bring normative rules to bear on the problem. However, as I wrote in chapter 3, it seems that normative performance on reasoning tasks is more to do with the ability to interpret the task as the experimenter intended, ignoring information that seems (but is not) relevant to the task as it is set. Summarising the

[186] Johnson-Laird casts this as a "distinction between *implicit* and *explicit* inferences [which] goes back at least to Pascal, and ... was revived by Johnson-Laird and Wason (1977)" (2004, p. 188, his italics).

[187] Care is needed in describing the commitments of such theorists. For example in a recent paper Evans (2006) finds a distinction in the literature between dual-process and dual-system accounts.

evidence, neurological as well as psychological, Noveck and Prado write that it shows that "correct performance on higher-level tasks has little to do with the better use of normative rules; it has more to do with avoiding biases while using such rules." (Noveck & Prado, 2007, p. 164). The evidence is just not there for the proposition that fast, unconscious reasoning is necessarily non-normative, i.e. that normative reasoning cannot be fast and unconscious.

I think the dual-process programme is in danger of conflating two distinctions. The first is the distinction between inference procedures and the associative links governing recall of information which feeds those procedures. The second is between explicit reasoning by value-preserving rules and reasoning by heuristics that take shortcuts[188]. In the recent dual-process literature, both of these distinctions have at times been mapped onto the distinction between slow, conscious and fast, unconscious processes. Keeping these distinctions separate is important for a clear view of heuristic processes.

In chapter 3 I have explained some of the forms taken by heuristics for reasoning. Heuristics may, by definition, reach inaccurate answers, and some heuristics proceed fast and automatically. However, some spontaneous processes are fully, canonically inferential, as when we rapidly deduce a proposition q from a sentence which expresses a proposition of the form $p \,\&\, q$, combine it with the already-known $q \rightarrow r$ and deduce r without awareness of making any of the steps. Conversely, some mere heuristics are consciously applied, as when we recognise that this is the Ruy-Lopez so we had better get our queen out early; or, trying to decide what to wear to a party, choose on the basis that 'you can't go wrong with a simple dress'. On the other hand, associative recall processes do seem to be inaccessible. One can attempt to 'jog' one's memory but, as the idiom implies, that is more like hitting the TV when it is on the blink than getting into the workings.

In making these points, I am simply advocating a now traditional account, according to which there is processing by computation over mental representations in short-term memory or memories, sometimes conscious, sometimes not. Whether these processes are conscious partly depends on what else is going on at the time. Playing chess or performing mental arithmetic might typic-

[188] Evans made essentially this point in a paper given at the *In Two Minds* conference, Cambridge, 2006.

ally be conscious activities (one heuristic, the other mostly canonical and value-preserving), but with enough practice they can be carried out subliminally, and not necessarily inaccurately, while carrying on a conversation, for example.[189] It is compatible with this view to suppose, in addition, that any great expenditure of effort by a procedure is very likely to come to conscious awareness, just as physical damage or unusual physical effort are brought to awareness through pain or discomfort.

The standard picture is that reasoning processes are fed by perception, linguistic parsing and associative recall of stored information. The processes involved in perception, parsing and recall seem to be inaccessible to consciousness. This does not amount to a dual-process account of reasoning, since these processes are not reasoning processes. One danger of dual-process theory, as of the importation into psychological science of the personal/sub-personal distinction[190], I think, is that conclusions may be drawn hastily about the nature of processes from the way that they appear to introspection.

My criticisms of dual-process theories extend to the views of some philosophers that judgements that are made by non-conscious processes are only to be explained in terms of neurology, rather than in terms of unconscious use of rules (e.g. Brown, 1988, p. 171ff). My reply to this point of view is to echo Rip's comments (above). Unconscious rules are well attested, in many areas of cognition, including reasoning.

[189] Cf Cherniak: "a person cannot, at one moment, think about all the information he possesses; he can only consider a subset of it. The contents of the short-term memory correspond to what he is now thinking about, not necessarily consciously (as when I drive a car properly while conversing about something else)" (Cherniak, 1986, p. 52). See also Sperber and Wilson (1986, p. 139) on the possibility that there is more than one short-term memory.

[190] I do not think that dual-process theory, even if successful, would support the view that the distinction between inferential and merely heuristic processes aligns with the philosopher's distinction between personal and sub-personal processes. Such a line has been taken, however (by the philosopher Keith Frankish in a talk at the 'In Two Minds' conference, Cambridge, 2006).

4.3.5 LOGICAL AND NON-LOGICAL PROCESSES

A view related to modern dual-process theories was held by Barnard (1968)[191], who divided processes for reaching judgements and choosing actions into logical and non-logical:

> By 'logical processes' I mean conscious thinking which could be expressed in words or by other symbols, that is, reasoning. By 'non-logical processes' I mean those not capable of being expressed in words or as reasoning, which are only made known by a judgement, decision or action. (Barnard, 1968, p. 302)

The similarities to modern dual-process theories of reasoning are marked: according to Barnard, non-logical processes are rapid and not explicit. There are also similarities to Recanati's views. According to both Barnard and Recanati there are essentially two types of mental processes: a) those involved in conscious thinking and reasoning; and b) processes which are non-conceptual, and therefore not consciously accessible, although they may produce conceptual output. Such views may be traced back to Thomas Reid[192], who defines reasoning as:

> the process by which we pass from one judgment to another which is the consequence of it. Accordingly, our judgments are distinguished into intuitive, which are not grounded upon any preceding judgment, and discursive, which are deduced from some preceding judgment by reasoning. (Reid, 1855, p. 423)[193]

Barnard's version of this two-process view is remarkably modern, given that his remarks predate modern cognitive science and the adoption of the symbol-system hypothesis. He writes that solving a quadratic equation fast uses

[191] Barnard's theory is set out in an appendix to Barnard, 1968, (originally published in 1938), based on a lecture given in 1936.

[192] Recanati cites Reid in this connection: e.g. Recanati, 2002b, p. 115.

[193] This statement of Reid's views is also compatible with my position (except in its use of 'deduced' where I would say 'inferred'). Recanati adds the additional assumption – which may be implicit in Reid's use of the terms 'intuitive' and 'discursive' – that reasoning is essentially a conscious activity. I leave the matter there, since exegesis of Reid's views is beyond the scope of this thesis.

"acquired knowledge ... marshalled and applied quickly." The person will be "unaware of what his brain actually does [and] unable to recall many of the broad steps that must have been taken." A human "could not write the text books that are registered in his mind" (Barnard, 1968, p. 306), as modern linguistics also teaches us.

We now know (or have strong reasons to believe, at least) that such systems as visual processing and language processing work in terms of the manipulation of symbols, but that we do not have conscious access to the workings of these systems – and that even if we did we might struggle to express them in words.

Inference and pseudo-inference

A view related to Barnard's is offered by Recanati. Recanati draws the dividing line between conscious and unconscious processes so that it coincides with Sperber's distinction between inference, including spontaneous inference, and non-inferential processing (Recanati, 2004, p. 43). Inference, as noted, relates input and output as premises and conclusion, and therefore both input and output of inference processes are, of necessity, conceptual representations.

Accordingly, formal processes which operate on non-conceptual representations are non-inferential. Processes of this type are sometimes described as inferential, however. According to Fodor, for whom all classical computations are by definition inferential (1983), the visual system performs a kind of encapsulated abductive inference. It takes visual stimuli as input. Its output is a representation of the scene that could have given rise to those visual stimuli. This process is "inverse optics": the mind has to solve the "problem of arriving at [representations of] surfaces from images" (Poggio, Torre, & Koch, 1985, p. 314), working backwards, as it were. The process can be seen as abductive 'inference' in the sense that the output explains the input, and the input underdetermines the output[194]: more than one scene could have given rise to the same visual stimuli, as many optical illusions demonstrate.[195] Recanati agrees

[194] There is more than one respect in which the visual system faces the problem that the input underdetermines the output: as well as the problem of inferring surfaces from images, there is the computation of 3D motion, again from two-dimensional cues (Poggio, Torre, & Koch, 1985, p. 314).

[195] Others have also claimed that visual processing is a form of inference, among them the sci-

with Sperber and Wilson that such processes, while they involve formal manipulation of symbols, are not strictly inferential. The key point, as noted by Sperber and Wilson (1987a, p. 737) (see also the discussion with references in Recanati, 2004, p. 41), is that the input to these processes is in the wrong form to perform inferences on. The input to the visual system is patches of light on the retina and corresponding activation of rods and cones in the eye, or perhaps patterns in the visual 'echoic' buffer, not propositions or proposition schemas. Activations in the visual buffer are non-conceptual both in the sense that they are non-propositional, and in that they are 'iconic' rather than 'discursive' (in the terms of Fodor, 2007). One cannot run *modus ponens* on an activation pattern or any other purely iconic representation.

I would call such processes pseudo-inference; Recanati refers to them as inference in the broad sense, distinguishing them from inference proper, which he calls narrow inference. Recanati's distinction is therefore able to accommodate cognitive science, while intended to support a binary division of mental processes into reasoning, which is essentially conscious, and non-logical processes, "not capable of being expressed in words or as reasoning".

R-availability

Recanati's view is more nuanced than the claim that only processes that are experienced as laborious and conscious at the time are inferential. In his view, narrow inferences may be made spontaneously or explicitly, but are characterised by their *availability* to consciousness[196]. When a narrow inference has taken place, the inferer is aware that she has made an inference, in at least the sense that she could (although she may not) bring to awareness all of: (1) the input to the inference, (2) the output, and (3) the fact that the input and output are inferentially related. This property (which I will call R-availability), is a necessary property of all narrow inference, according to Recanati, and of reasoning.[197] Recanati discusses Sperber's example of spontaneous inference.

entist Hermann von Helmholtz, who wrote of perceptions that "...by their peculiar nature they may be classed as conclusions, inductive conclusions unconsciously formed" (von Helmholtz, 1962) cited by Barlow, 2002, p. 602 (although Kubovy & Epstein, 2002, p. 619 claim that for Helmholtz the kind of inference involved was deductive).

[196] Recanati's availability principle as it relates to 'what is said' was stated in Recanati, 1989.

[197] I am hedging here because Recanati does not clearly distinguish reasoning from narrow

Hearing the doorbell ringing, you form the belief that there is someone at the door. For Sperber, (and I agree) this example illustrates that beliefs can be reached without "deliberate, conscious inference", but still inferentially. Recanati, on the other hand, claims that this sort of inference *is* conscious, since it is R-available. The input to the inference is the belief (itself acquired directly from perception) that the doorbell is ringing. This and the output belief are both "conscious and available to the subject" (Recanati, 2004, p. 42) according to Recanati, and the subject is also (potentially) aware that the output is inferentially grounded in the input: asked how she knows there is someone at the door she could reply: "Because I heard the doorbell."

R-availability does not imply that the inferential steps made, or even the type of inference involved, can be brought to consciousness. That would be contrary to the evidence. As O'Brien writes:

> we have no reason to expect that ordinary reasoners would monitor the sorts of processes they use to obtain any particular inference; that is, ordinarily people are not aware of whether an inference stemmed from logical, pragmatic, or any other sort of inference-making process, including from general epistemic knowledge, but would know at most that some proposition has been inferred. (O'Brien, 2004, p. 210)

That this is a fair point might be illustrated by the considerable argument, among experts in the field, about which abilities are tapped by the selection task. Apparently we do not have reliable intuitions about whether our inferences are analytic or pragmatic, and whether the principles used are domain-general or domain-specific or due to use of the faculty for utterance interpretation.

The personal/sub-personal distinction is also central to Recanati's distinction. For him all reasoning and narrow inference is personal; and no sub-personal process counts as reasoning or narrow inference, even if the processes causally involved are parallel to conscious reasoning. Recanati distinguishes between tacit sub-personal inferences and tacit personal inferences (the distinction is due to Garcia-Carpintero, 2001). Tacit sub-personal inferences are

inference in his comments on this subject. As noted, he cites Reid's definition of reasoning, but prefers the term (narrow) inference in exposition of his own views.

inferences only in the broad sense: they are those which are "ascribed to a cognitive system merely on the grounds that 'the causal processes constituting the system mirror the processes of someone who [performed] the relevant [inferences] in an explicit form'" (Recanati, 2004, p. 49). For an inference to count as a tacit personal (narrow) inference, the rational agent who makes it must be "capable of making the inference explicitly and of rationally justifying whatever methods it spontaneously uses in arriving at the 'conclusion.'" (Recanati, 2004, p. 49) The spirit of these claims, I think, is more conceptual than empirical. In other words, I think Recanati takes these remarks to amount to clarification of the concepts *personal*, *sub-personal* and, especially, *inference* and their relations to each other.

Recanati's views on inference are the key to understanding his view of pragmatics. Spontaneously drawn narrow (i.e. personal) inferences provide Recanati with an answer to the question of how Gricean pragmatics is explanatory. On this view, the explanatory power of Gricean-type derivations in pragmatics is that reasoning/narrow inference is R-available: it can be brought to consciousness at the time or after the fact. Thus whenever there is narrow inference (spontaneous or otherwise) in pragmatics there is an awareness of the input, the output and the fact that they are inferentially linked: the essentials of a Gricean explanation.

So in implicature derivation, which Recanati says is narrowly inferential, the idea is that the hearer must be able to be consciously aware of the derivation of implicatures: more specifically, that he must be capable of being aware of the input to the derivation, the fact that some proposition *p* (what is said) has been expressed, and of the output, the implicature or implicatures, and of the fact that the implicature is the output from a personal, R-available inference process with *p* in the input.

I share with Recanati the view that there is spontaneous pragmatic processing carried out by reasoning/inference: in Recanati's theory, for implicatures; in my opinion, for interpretations as a whole. On the other hand, there are some marked differences between my view and his. The first is Recanati's claim that narrow inference and reasoning are necessarily consciously available. Secondly, Recanati divides pragmatic processes into primary and secondary. The third difference is that given this division, I think that there are worries about the explanatory adequacy of Recanati's account of primary processes, and also, for different reasons, secondary processes.

Primary and secondary pragmatic processes

Recanati, like relevance theorists (e.g. Wilson & Sperber, 1981; Carston, 2002b) and other radical pragmatists, stresses that the proposition expressed by an utterance is considerably underdetermined by the linguistic facts about the utterance, so that there is a need for considerable pragmatic processing to get to the explicit meaning of an utterance from the kind of representation that would result from processing according to rules or principles in the grammar. However, he differs sharply from relevance theory in proposing that explicit meaning and implicit meaning are arrived at by two distinct types of mental activity, only one of which is properly inferential.

For Recanati, primary processes are non-inferential and sub-personal; secondary processes are inferential and personal. Secondary processes are the usual Gricean inferential derivations of implicatures from what is said, (or the fact that it was said, or the manner in which it was said) and given that they are personal, narrowly inferential processes are R-available. Primary processes, which derive what is said, in a propositional form, from the linguistic input, are non-(narrowly) inferential and not R-available, according to Recanati.

The division into primary and secondary processes could be seen as an echo of Grice, since Grice only discussed the use of the Cooperative Principle and maxims in arriving at implicatures, leaving open the question of what principles govern processes such as reference assignment and disambiguation which contribute to what is said. However, there is a crucial difference: as discussed above, Grice thought that recognition of the intention behind an utterance provides a *reason* for the hearer to think that the speaker believes a particular proposition (or wants him to entertain this proposition, or to have a certain other response), and that the intention is 'recognised', or rather worked out on the basis of the normal meanings of the expressions used and the context. The hearer has reasons for entertaining, and arrives by reasoning at, the explicit meaning of an utterance – what is said – as well as at implicatures of the utterance. Thus, although Grice discusses calculability only for implicatures, potentially there is a Gricean story about arriving inferentially at explicit meaning too. In contrast, Recanati presents a picture of the derivation of what is said as a clearly non-inferential process, determined by brute facts about accessibility of senses of words and of referents[198]. (See Recanati, 2004, p. 30, for example derivations.)

> ... the interpretation [of the explicit meaning of an utterance] which eventually emerges and incorporates the output of various pragmatic processes results from a blind, mechanical process, involving no reflection on the interpreter's part. The dynamics of accessibility does everything, and no 'inference' is required. In particular, there is no need to consider the speaker's beliefs and intentions. (Recanati, 2004, p. 32)

I have not taken it as one of the main tasks of this thesis to argue for an inferential view of pragmatics against such opponents as Millikan and Burge. My aim is to argue that a broadly Gricean inferential-intentional approach is compatible with a realistic view of rationality, and to explore the consequences of the combination. The fundamentals of the Gricean approach are mostly presupposed by this endeavour. Similarly, I do not think that it is a central concern of this thesis to argue, contra Recanati, that it is utterance interpretation as a whole that is inferential, rather than the derivation of implicit meaning only. I think, however, that there are reasons to oppose the claim that the pragmatic processes involved in reaching explicit meaning are non-inferential. I have tried to show that Grice's account of speaker meaning rests on the causal and justificatory force of speaker intentions, so that Gricean accounts of how utterances as a whole are understood, as well as Gricean accounts of implicature derivation, appeal to reasoning about speaker intentions as an answer to both the *how* and the *why* questions.

We might also challenge the claim that the pragmatic processes involved in reaching explicit meaning are unavailable. If we ask someone why they think that speaker S meant *p* (where *p* is the proposition expressed by an utterance) they might well say something like, "Because I heard her say *x*", or (if schooled in modern linguistics and the work of Grice) "Because I heard her

[198] Recanati's account of what is said still has an inferential flavour, however, in that what is said is represented as speaker's meaning, as I commented in chapter 1.

say *x* and I had no reason to think that she was speaking ironically or otherwise didn't mean what she said, and the form of words used in *x* clearly conveys *p*, due to the meanings of the words and the syntactic structure, or so syntacticians and semanticists tell me." For aspects of explicit meaning, similar points apply: e.g. "I knew you were talking about Recanati because you kept pointing at him"; "I thought that you meant *bank* bank, not *river* bank, because I asked you if you had any money, so why would you start talking about river banks?"

The point of this objection is that the hearer seems to have perfectly good reasons for thinking that the speaker meant *p*, may well be aware of these reasons and of the fact that they are reasons for thinking that S meant *p*, and may even be able to state them (with more or less precision, no doubt, depending on how thoughtful they are, how diligently they read Grice, and other factors).

Robyn Carston has raised essentially this objection to Recanati's claim that primary pragmatic processes are not R-available:

> ... surely most hearers are able to perform the reflective activity of 'making explicit' their tacit reference fixing process: if asked how he knows that the speaker was referring to Tony Blair (rather than Cherie Blair or John Prescott), the addressee could respond that he knows this because the speaker used the word "he" while pointing at (or demonstrating in some other ostensive way) Tony Blair. He thereby shows that his referential hypothesis has a rational basis and that he is consciously aware of both the hypothesis itself, the evidence on which it is based and the relation (inferential?) between them, and that, on reflection, he is able to make the connection explicit. (Carston, 2003, pp. 1– 2)

I do not think that Recanati would want to deny that hearers are capable of offering rationalisations for some explicit parts of speaker meaning. How can this fact be made compatible with Recanati's claim that primary pragmatic processes are non-inferential and un(R-)available? I think Recanati has to say that it is not simply the R-availability of a narrow inference that marks out a process as inferential, but that R-availability is an essential property of that (type of) process. In fact, in a section that is explicitly a reply to some of Carston's criticisms, Recanati writes:

> It is constitutive of conversational implicatures that the inference that gives rise to them is available to the interpreters ... On the other hand, I maintain that the reflective capacity to rationally justify one's interpretation is not constitutive when the interpretation involves only primary pragmatic processes. (Recanati, 2004, p. 50)

If I am right, Recanati is saying that secondary pragmatic processes are essentially R-available and primary pragmatic processes are essentially un(R-)available and non-inferential (in the narrow, proper sense of inference). That is, his response to Carston is that you can sometimes consciously construct a kind of inference that could have led from an utterance to what is said, but this is *only* a rationalization (in the usual sense: an explanation in terms of reasons that may not have been operative) and is not enough to show that explicit meaning is arrived at inferentially.

Recall that Recanati makes use of the personal/sub-personal distinction, distinguishing between tacit sub-personal inferences and tacit personal inferences. Recanati says that for implicatures:

> A tacit inference is ok, provided it is of the 'personal' sort, i.e. provided the subject herself has the reflective capacities for making the inference explicit. To say that this capacity is constitutive, in the case of conversational implicatures, is to say that there would be no conversational implicature if the interpreters did not have that reflective capacity. (Recanati, 2004, p. 50)[199]

This raises the theoretical question about reasoning: Is being able to reflect on an inference really determined by what kind of inference it is? This question ties in with a worry that may be empirical or conceptual. Recanati can be taken as making the empirical prediction that people who cannot consciously reason about intentions cannot derive implicatures. Alternatively he may be making the conceptual point that the concept of implicature should not apply

[199] This quotation is extracted from the middle of the immediately previous quotation, i.e. in the original text this section fills the position of the ellipsis in the quotation above.

to any mental representation that such an agent might derive from an utterance.

As we have seen in chapter 1, there is developmental evidence against the empirical prediction: very young children apparently lack key elements of the ability to reflect on mental states such as beliefs and desires, but they comprehend some implicatures and other pragmatic phenomena. As I have discussed, a number of studies show that children fail false belief tasks until around four years old (e.g. Wimmer & Perner, 1983; Clements & Perner, 1984; Perner & Lopez, 1997; Templeton & Wilcox, 2000). Children are capable of pragmatic interpretation much earlier than this. Recent work by Pouscoulous and Noveck (2004) shows that even the youngest children tested (around 4 y.o.) are capable of implicature retrieval if the cognitive demands made by the experimental task are low enough, as Noveck (2001) anticipated. Developmentally, it seems that the ability to derive implicatures precedes general reflective reasoning about the beliefs of other agents. If Recanati's claim is to be taken as an empirical prediction, there is growing evidence against it.

If, on the other hand, Recanati's point is conceptual rather than empirical, it seems that he is committed to the claim that when young children – who cannot consciously, reflectively reason about intentions – understand utterances, including implicit elements of the meaning, we cannot regard what has happened as involving implicature derivation. It would be strange to say this if it turns out that young children understand a speaker's implicated meaning by identical mechanisms to adults and arrive at the same mental representations in particular cases. These children would be making the same inferences according to the same causal processes as adults. The only difference is that the adults are able, after the fact, to bring some aspects of the inference process to conscious awareness. If this is what Recanati means, then I think that his suggestion should be rejected. As theorists we can decide how to define the term *implicature* and how to refine its definition in the light of evidence[200]; I see no point in defining it so that identical inferences carried out by essentially the

[200] It seems that the meanings of theoretical terms are liable to change as understanding deepens, whether they are terms originally from natural language, such as 'force', or were always terms of art, like 'implicature'. See Reid, 2004, pp. 53-54, on 'number', 'multiply' and 'divide'.

same causal processes sometimes do and sometimes do not count as implicatures, depending on some other ability of the agent.

Recanati's views of pragmatics no doubt deserve more thorough discussion than I have the space for here. The discussion in this section is simply intended to suggest some of the problems that arise for theories which try to take reasoning or (narrow) inference out of pragmatic interpretation (or out of part of it) and for theories which claim that pragmatic inference is essentially personal, conscious and available. I reserve for the next section, which concerns the specifics of the heuristics involved in utterance interpretation, two further worries about Recanati's views: that he offers no causal account of the processes involved in spontaneous implicature derivation, and that his account of the derivation of explicit meaning is implausible precisely because the processes proposed are not inferential, so that in his picture there is no element of trial-and-error problem solving.

4.3.6 CONCLUSION

The aim of this section has been to advocate a particular way of looking at the role of rationality and reasoning in pragmatics. A secondary claim is that the view I have outlined was essentially Grice's view. I have argued for a controversial reading of Grice, suggesting that he saw the retrieval of implicatures as a case of reasoning. This seems to me to follow naturally from comparison of what Grice says about the hard way and the quick way of reasoning with his insistence on calculability of conversational implicatures. More broadly (and less controversially) meaning and reasons are intimately related in Grice's work. Recognition of the intention behind what is uttered and how it is uttered not only causes the hearer to entertain an interpretation of the utterance, but also provides the hearer with reason(s) for that interpretation, at least in the 'anaemic' sense standard in causalist theories (that there seem at the time to the hearer to be such reasons). It is traditional to see working with reasons as reasoning, and as Grice suggested, a picture of reasoning as the construction of trains of thought characterised by transitions that preserve rational acceptability shows how this can be so. Thus there is a Gricean account in terms of reasoning of the interpretation of speaker meaning.

While this section has concerned the role of reasoning and inference in pragmatics, it has also been necessary to comment on dual-process theories and related views according to which fast, subliminal processes are necessarily merely heuristic. I agree with Simon[201], Cherniak[202] and Sperber and Wilson that the main task in the study of inferences or reasoning in cognitive science is to investigate the kind of inferences made and the mental mechanisms involved in making them. However, as I have shown, a number of theorists – in pragmatics, and in reasoning – have thought that whether a process is conscious, or available to consciousness, tells us something important about its status. So part of the burden of this section has been to explore representative versions of these views and suggest that it is not necessary to hold such a view to see broadly Gricean explanations in pragmatics as explanatory.

The picture that emerges is that pragmatic interpretation is carried out by goal-directed inference, regardless of whether the inference is conscious or not, consciously available or not, personal or sub-personal. As we have seen, this picture is incompatible with the views of some theorists. Some, like Campbell, think that only conscious processes are inferential or count as reasoning. For them, when pragmatic interpretation is unconscious it must be using different mental processes: mere heuristics. Recanati's nuanced view is that the only inferential pragmatic processes are those which are personal and can be made conscious. In contrast to both of these views I have argued that whether a pragmatic process is conscious, or can be made conscious, tells us nothing *in principle* about the kind of process it is. I assume that the blend of heuristics and canonical warrant-preserving transitions involved in pragmatic interpretation is largely a matter for empirical investigation. As in study of

[201] Simon wrote in 1997 that in 1946 he had:

> finessed the issue by assuming that both these processes [conscious and subconscious] were essentially the same: that they draw on factual premises and value premises, and operate on them to form conclusions that became the decisions (Simon, 1997, p. 131).

Simon of course, presents much evidence for unconscious rational activity, e.g.:

> It has been shown that many of the steps in mathematical invention – than which there can presumably be nothing more rational – are subconscious; and this is certainly true of the simpler processes of equation solving. (Simon, 1997, p. 84)

[202] "This [i.e. Cherniak's] rationality theory continues to be significantly idealised ... I will ... not distinguish between deliberate conscious inference and unconscious inference." (Cherniak, 1986, p. 5)

natural language syntax, introspective evidence plays an important role, but we should not expect that we have reliable intuitions about the processes or principles involved[203]. Our intuitions are primarily about the felicity and immediate implications of the interpretation: not the process but the product[204], although in pragmatics plausible rational reconstructions can generally be made because the input and output are both conceptual.[205]

I leave undetermined one question about Grice's views in this area. I do not know whether Grice would have endorsed Recanati's position that only processes that the inferrer himself can become aware of as inferences are in fact inferences. As I have said, my own view is that this position is theoretically unstable and empirically untenable in the face of evidence that pragmatic inference can be carried out by children too young to reason consciously and explicitly about beliefs and intentions.

In the course of this chapter (and in chapter 3) I have given examples of heuristics that are consciously applied, and of unconscious processes that are algorithmic and warrant-preserving. I have little doubt, however, that heuristics are central to pragmatic processing and that most pragmatic processing is carried out 'beneath' conscious awareness. In the final chapter I consider the specifics of pragmatic processing in more detail.

[203] Noveck and Sperber (2007) say that pragmatic intuitions are much less direct than semantic intuitions, since they involve reflection on imagined utterances in constructed scenarios, as syntactic and semantic intuitions (presumably) do not.

[204] Nicolle and Clark (1999) found that when participants were asked to paraphrase what the speaker has said, in cases where there was one strong implicature this was often given; in other cases the proposition expressed was paraphrased. One implication of this research is that the product of utterance interpretation cannot always be picked apart by the hearer into what is said and what is implied, even reflectively, after the fact: an uncomfortable result for Recanati's availability principle, as Nicolle and Clark suggest. Indeed, if the availability principle is taken seriously, these results should lead to the conclusion that it is the input to pragmatic processing and the output taken as a whole that are available, supporting the theory that pragmatic processing has one phase rather than two.

[205] N.B. This is quite different from Recanati's view. My view is that (normal, adult) humans can reason with conceptual representations, so we can generally reconstruct any inference. An indication of the difference is that my view does not entail, where Recanati's does, that an agent is able necessarily to reconstruct his own inferences.

Chapter 5 · Conclusion: The comprehension heuristic

> a principle that is implied in all rational behavior [is] the criterion of efficiency... to be efficient simply means to take the shortest path, the cheapest means, towards the attainment of the desired goals (Simon, 1947, p. 12)

5.1 INTRODUCTION

In the previous chapter I have explored connections between rationality and communication, particularly in the interpretation of utterances, arguing that reasoning plays a crucial role. I have suggested, given that human rationality is bounded, that much of this reasoning involves shortcuts, many of them heuristics in the sense discussed in chapter 3: non-algorithmic procedures which do not guarantee reaching the right answer. In this final chapter I look in more detail at the processes for working out what a speaker meant by her utterance. I assume that the fine detail of the processes involved in the inferential recovery of speaker meaning is largely an empirical matter. My comments here will be more general in nature, describing some properties that I would expect the processes to possess, given the task they face and the limits on human rationality.

The problem of utterance interpretation is a rather ill-structured problem in Simon's terms. I have claimed in chapter 3 above (following Simon) that aspiration-level, sequential search heuristics, guided by recognition of features of the problem, are a solution to the general problem of how cognition deals with ill-structured problems, including problems of abductive inference. Sperber and Wilson's work on comprehension is a rich source of ideas about how an inferential utterance-interpretation procedure can be boundedly rational. The relevance-theoretic comprehension procedure that they propose has some interesting properties tailored for the domain of intentional-inferential pragmatics, taking advantage of certain ways in which the task of comprehension is not completely ill-structured.

I look at several properties of the comprehension heuristic, including the need for aspiration-level search, the least-effort path followed and other constraints on the solutions generated, the role of feature-driven recognition, the stopping rule used and the overall frugality of the procedure. This last point leads to consideration of dedicated heuristics as a solution to Fodor's problem. This solution differs from the strong modularity possessed by peripheral processes. I comment on the way that encapsulation in terms of the process used, together with the frugality of that process, ensure that Fodor's problem does not arise in practice for utterance interpretation. The final contention of this thesis, then, is that a dedicated fast and frugal heuristic is a cognitively realistic, boundedly rational implementation of inferential-intentional pragmatics.

5.2 INFERENCE AND LOW-LEVEL EXPLANATION

In the previous chapter I have spent some time describing the views of François Recanati. I raised but left unanswered one question about the difference between his views and mine. I think that there are worries about the explanatory adequacy of both Recanati's account of primary processes, and, for different reasons, his account of secondary processes. I discuss these points here because I think that they illustrate, on the one hand, why there must be a low-level account of the processes involved in comprehension, and on the other, that the processes need to be inferential.

The simpler of the two points is the first, which I believe applies to Recanati's account of 'secondary' processing: implicature derivation. Recanati says – and I agree – that implicature derivation is inferential. If one accepts, as I do, that the steps involved in such inferences are (typically) not isomorphic to the steps in valid arguments then the question arises of how it is causally explanatory to claim that utterance interpretation, or some part of it, is inferential, or an instance of reasoning. This is the modified version of Warner's question in section 4.3.4 above.

My answer is that it is not enough simply to claim that utterance interpretation is reasoning, or that it is inferential. That claim ties together the answers to the *how* and the *why* questions discussed above, but it does not fully answer the *how* question, because it says only that in some sense the process is like a fully warrant-preserving derivation. A scientific pragmatic theory must attempt an account of the heuristic processes involved in utterance interpretation – and explain in what way such processes are inferential. The processes that accomplish reasoning are often faster and experienced as more intuitive than fully explicit inferential derivation. An account of pragmatic processing (or part of it) that claims only that utterance interpretation is a kind of reasoning which happens quickly through heuristic processes might be satisfactory as philosophy but not as a full scientific explanation. What is needed in pragmatics is an account of such processes.

When we consider the low-level account of utterance processing that is needed[206], we can see that there should be consequences for the kind of account postulated that we are concerned with heuristics that perform *inference*. This is the second point that I want to make by comparison with Recanati's account. In particular, for the comprehension process to be inferential rather than 'blind' or 'brute', there must be, in principle, some evaluation of solutions, rather than simple generation of one solution.

5.2.1 SEQUENTIAL TRIAL-AND ERROR SEARCH

Given that utterance interpretation requires rapid choice of a good solution from an open-ended set of alternatives, there is no realistic alternative to sequential trial-and-error search, for reasons discussed in general terms in chapter 3. It costs time and effort to search. Optimising theories (in Simon's sense) simply assume that the best available solution will be found, as though all the alternatives were generated, evaluated and compared. Theories that idealise away from search in this way are unrealistic. As I have explained, it also costs to calculate the potential benefits of search, and the costs of this kind of calculation are generally prohibitive, so theories which posit optimisation under constraints are also unrealistic. Therefore, if the solution to a problem is to be picked from among a set that is not known in advance to be

[206] Such accounts, are, of course, still somewhat idealised, as noted in chapter 3. The way that sequential trial-and-error search is implemented neurologically is a separate question. I have said that facts about accessibility help determine which solutions are tried. This could be seen as part of an underlying 'pandemonium' account of cognition, with competition between processes, and parts of processes, for resources.

limited, then we cannot assume either that it is 'as if' all possible solutions were considered, or that the costs and benefits of continuing the search were calculated at each point in the search.

Sperber and Wilson make a further point about processes in which all potential solutions are generated or found and ranked. If this must always be done, then every search will consume the same amount of effort: the effort required to generate and compare all alternatives. If the number of alternatives is large, or the cost of generating them is high, then the effort required to generate all of them will be prohibitive. In the case of communication this would mean that it would never be worth the effort of processing an utterance: "It is hard to think of any ostensive stimulus that would be worth such an absurd amount of effort." (Sperber & Wilson, 1986, p. 166)

Therefore I proceed on the assumption that comprehension is carried out by a process that generates solutions one by one and tests them for acceptability. This comprehension procedure must have a stopping rule that is well suited to the problem domain so that it tends to stop search when continuing would not be worthwhile. It must do so without exhaustive calculation of the pay-off that would be expected if the search were continued.[207] Such a heuristic is well-suited to implement inference because it tests solutions and can reject them. To demonstrate what I mean by this, I compare what Recanati says about primary processes with the kind of heuristic that I suggest.

5.2.2 ACCESSIBILITY-ONLY ACCOUNTS

Recall that Recanati's claim is that all the work for arriving at the explicit meaning of an utterance ('what is said' in his terms), is done by salience and accessibility. Recanati says that this is "a blind, mechanical process... The dynamics of accessibility does everything, and no 'inference' is required." (Recanati, 2004, p. 32) The essential difference between this kind of account and an inferential account is not necessarily in the intricacy of the processing

[207] Sperber and Wilson made a related point about the strength of assumptions. They hypothesise that the mind does not generally represent the degree of likelihood with which each belief is held. Accordingly, there can in general be no calculation of the likelihood of a belief on the basis of the likelihoods of the assumptions which support it, nor can there be comparisons of degrees of confidence across domains (Sperber & Wilson, 1986, pp. 75–83).

involved, nor in the amount of information brought to bear. I have said in chapter 3 that for complex problems 'recognition' (in Simon's sense) of features of the situation brings into play memory, itself structured in terms of frames and schemas. This is so for an inferential account such as Sperber and Wilson's (1986, p. 137ff), as for an accessibility-only account. The difference between an inferential account and an account like Recanati's is partly in constraints on the solutions generated in an inferential account (which I consider below), and partly that for Recanati's 'blind' process, the solution reached is the only solution, whereas for a trial-and-error search process, any solution generated is only a potential solution until it is evaluated and accepted.

In either account, accessibility will determine the first interpretation reached. Consider for convenience the limited problem of determining the referent for the pronoun 'he' in sentences (21) a–c, neglecting any possible referents other than the individuals picked out by the DPs 'a policeman' and 'John'. (See Recanati, 2004, p. 32.) Various factors compete in influencing the accessibility of referents. The subject of a sentence is prominent, so the referent of 'a policeman' has an advantage, but 'John' is closer to 'he' than the subject, so John may be a more accessible referent on that count. In (21a), the decisive factor appears to be the predicate "steal a wallet", which may raise the accessibility of a stereotype or frame in which policemen attempt to catch criminals, thus making John the more accessible candidate for the remaining role in the frame, the culprit. In a similar way – although the knowledge involved is less stereotypical – the policeman may be the more accessible referent in (21b).

(21) a) A policeman arrested John yesterday; he had just stolen a wallet.

b) A policeman arrested John yesterday; he had needed one more arrest to qualify for the end-of-year bonus.

c) A policeman arrested John yesterday; he had just taken a bribe.

The results on a particular occasion may depend on activation that is due to ideas that are 'in the air'. If the remark follows conversation about a recent crackdown on corruption, then John might be the more accessible referent in

(21c) – unless the crackdown was on corruption in the police. The difference between a model like Recanati's and a trial-and-error search model of inference is that in the inferential model the most accessible interpretation is assessed and only then accepted as the speaker meaning. In Recanati's model the interpretation that is most accessible when the point of closure is reached is taken as the explicit meaning of the utterance.

Sperber has raised essentially this point about Recanati's model (Recanati, 2004, p. 32). His objection is that there seem to be cases where pragmatic processing is subject to garden paths: processing goes in the wrong direction before it finds the right solution. However, as Recanati points out, a 'blind' process can appear to exhibit similar behaviour if it receives input over some period of time. As the words of a sentence come in, they will raise the accessibility of certain frames or schemas in long-term memory. The process can be seen as occurring in a network of propagating activation, where the values at the inputs are set one by one and the solution is (or is read off, or otherwise determined by) the final state of the network. There can be transient states of the network on the way to its final state that are quite different from the final state, since a word that comes late in the utterance may considerably affect the final state reached. A further possibility (not mentioned by Recanati) is that garden path effects might arise in a network of this type because of the finite speed of propagation of activation through the network. In effect the network might be a bit 'springy' and would then take time to settle down to its final state, even neglecting the fact that information arrives sequentially in comprehension.

I agree, then, with Recanati that the simple possibility of garden-path effects is not in contradiction with a non-inferential account of pragmatic processing (or part of it) like his. However the point that I want to make is illustrated by the contrasting positions taken by Recanati and Sperber. In an inferential model like Sperber and Wilson's it is always a possibility to reject an interpretation. In a 'blind' model it is not: the bare facts about accessibility and the structure of memory, including frames and schemas, have to make the interpretation come out correctly. A theory of this sort is a bet that except in cases where the speaker is misunderstood, our memories are structured so that the first interpretation that comes to mind – at closure of the process, once the sentence has been parsed in the linguistic and non-linguistic context in which it is uttered – will always be right .

This is a very strong claim, for which strong evidence would be needed, in my opinion. I do not go into the evidence here. Wilson and Matsui (1998) discuss some shortcomings of accessibility-only accounts. As they write, "most work on reference resolution... acknowledges,... that the most accessible candidate can be rejected and another selected on pragmatic grounds." (p. 177)

A slightly more general claim is that a pseudo-inferential account of utterance interpretation can work. Is it possible to explain the derivation of explicit meaning of utterances, including reference assignment, disambiguation and lexical enrichment, solely in terms of – perhaps very intricate – routines like the ones in visual processing for edge detection? This is an empirical question. In my opinion, the evidence is that it cannot be done this way because of the interdependence of explicit meaning with implicit meaning (which has to be inferred), as I discuss below. Even frameworks that have tried to bite off only a part of the problem have run into serious trouble because of this kind of interdependence. An example is Discourse Representation Theory, which has these difficulties despite the assumption, made from the start, that some part of the account – of the determination of referent of a pronoun, for example – must lie outside the system. Hans Kamp's original interpretation rules require the selection of a "suitable" referent for the pronoun, acknowledged to be a "deliberately 'fudgey' formulation" since "To state what ... the set of suitable referents is, we would have to make explicit what the strategies are that speakers follow when they select the antecedents of anaphoric pronouns" (Kamp, 2002, p. 215), that is, to do full-blown inferential-intentional pragmatics. Even an approach as aware of pragmatics as Kamp's faces the question of whether it is worth proposing complex bottom-up algorithms for (e.g.) reference resolution when it is clear that in general such questions ultimately depend on hearer inference about speaker intentions. (Breheny, 2003 presents the case for simple semantics together with inferential pragmatics as opposed to complex 'dynamic' algorithms in semantics.)

It is worth noting that the claim made by an accessibility-only account like Recanati's is stronger than a related claim that needs to be made about sequential search if such processing is to be frugal. In frugal sequential search, the correct solution is reached very quickly. The weaker claim is that accessibility factors are *partly* responsible for this rapid zeroing in. It is a weaker assumption that our minds are well-enough attuned to the environment for accessibility to *help* reach the correct solution quickly, than it is to assume that such accessibility factors *always*[208] deliver the correct solution.

5.2.3 CONSTRAINTS ON SOLUTIONS, AND THE STOPPING RULE

There are at least two other factors involved in helping to ensure that the comprehension heuristic rapidly arrives at the correct solution. I have discussed such factors in general terms in chapter 3, where I have noted that constraints on solutions generated are defined by the kind of problem addressed. The first such constraint on the solutions accessed that I want to discuss is Sperber and Wilson's claim that the search follows a least-effort path. This depends on a property that is specific to the comprehension process. In general, the only justification for following a least-effort path in search is if there is good reason to expect a good solution to lie on that path. Many searches that we conduct involve considerably more than least possible effort, including searches that involve generation of solutions. For example, academics (mostly) do not take a least-effort approach to research, writing the sloppiest possible version of a paper, only rewriting if and when it is rejected and then only in minimal ways.

Sperber and Wilson have argued that in comprehension it is reasonable to assume that the correct solution is on a least-effort path, because each utterance carries a presumption of optimal relevance. (In technical terms, the speaker makes this presumption manifest in making an ostensive act). The presumption was given in (2), repeated here:

[208] Cases of miscommunication aside, as noted above.

(2 repeated) *Presumption of optimal relevance*

a) The ostensive stimulus is relevant enough for it to be worth the addressee's effort to process it.

b) The ostensive stimulus is the most relevant one compatible with the speaker's abilities and preferences. (Sperber & Wilson, 1995, p. 270)

The hearer has no option but to assume that the speaker is rational, albeit boundedly so, since if the speaker is not rational at all there is no reason to suppose her actions serve her intentions. Therefore, as Sperber and Wilson say, "a rational communicator, who intends to make the presumption of relevance manifest to the addressee, must expect the processing of the stimulus to confirm it" (Sperber & Wilson, 1986, p. 165). Thus the hearer can proceed according to the assumption that the speaker will not have put him to gratuitous effort: the effort must at least merit the pay-off. This means that the correct interpretation should be on the least-effort path, since otherwise an interpretation that was on the least-effort path, but not the intended one, might stop search before the intended one is ever reached. Another way of seeing this point is to consider an intended interpretation conveyed using a rather obscure stimulus, so that it is off the least-effort path. If the hearer finds the intended interpretation at all, then it will have required considerable effort. The stimulus will therefore not satisfy clause (b) of the presumption, given that relevance is lower when effort is higher, and that some other stimulus, requiring lower effort, could have been used. Equally, since the presumption is symmetrical in efforts and effects, an interpretation that requires little effort but delivers inadequate effects for that effort will be rejected. (See discussion at Wilson & Sperber, 2002, p. 605ff. There is related discussion at Sperber & Wilson, 1986, pp. 168–169, based on an older version of the presumption of relevance.) These are very strong constraints on the alternatives that need to be generated and tested.[209]

[209] As Wilson and Sperber note, there are complications connected with the strategies employed by sophisticated hearers who know that a) speakers may be mistaken about the relevance of their utterance to the hearer, and b) speakers may deceitfully produce utterances that are only intended to *seem* relevant (Wilson & Sperber, 2002, p. 605 note 5). For discussion see Sperber, 1994; Wilson, 2000.

In fact the intended interpretation must either be the first one reached by following the least-effort path that satisfies clause (a) of the presumption of optimal relevance, giving an acceptable balance of cognitive effects for the effort put in, or there must be some reason to suppose that a more relevant stimulus can be found further along the least-effort path. In cases where there is reason to think that a more relevant stimulus can be found, clause (b) of the definition is not satisfied, and this mandates further search. For reasons already discussed, the determination of whether a particular interpretation satisfies clause (b) cannot be made by exhaustive calculation of the costs and benefits to be derived from further search: the decision about whether to continue must be available without such calculation. It could be that the search is currently looking very promising, in that it is bringing large returns for little effort, so that further solutions are, on the face of it, worth considering. Or there may be a specific expectation in the context which makes it clear that the interpretation under consideration is not as relevant as the speaker must have intended her utterance to be. These criteria are not mutually exclusive.

Note that what counts as a good return is dependent not only on expectations, but on what returns are being derived by other cognitive processes with which limited resources must be shared. I would expect that for fast, largely automatic, central processes the situation is similar to what has been found in the psychology of attention for perceptual processing. The work of Lavie and colleagues has demonstrated that in perception (auditory as well as visual) the depth of processing of any stimulus (or more accurately 'channel' of stimuli: e.g letters appearing at the top-right of a screen) depends on how much effort is being expended elsewhere. If the channel on which attention is mainly focussed requires little effort, then other channels are simultaneously monitored and processed. Conversely, if the attended channel requires a high level of effort, stimuli on other channels are not processed to any depth (Lavie & Tsal, 1994; Lavie, 2000; Lavie, 2006).[210] In general, it would be surprising if cognit-

In principle there are two ways that these complications might be accommodated in the current framework: (1) in terms of the sophistication of the expectations of relevance brought to comprehension by the hearer, with search following a least-effort path but stopping at a different point (Sperber, 1994); or (2) in terms of perturbations from the least-effort path.

[210] Pashler (1998) reaches a similar conclusion from an extensive review of the literature (but

ive resources were not allocated to processes on the basis of expected returns, modulated by attentional factors[211].

Returning to utterance interpretation, the implication of the preceding discussion is that the presumption of optimal relevance both constrains the solutions generated by showing that they must lie on a least-effort path and mandates a two-criterion stopping rule: stop if both a) the interpretation being tested is worth the effort expended in the search so far (given expectations); and b) there is no indication that a solution with a better balance of pay-off to effort put in (and with, necessarily, therefore, a higher pay-off) can be reached by continuing.

Sperber and Wilson give an example where the second part of the stopping rule comes into play:

(22) Henry: Do all, or at least some, of your neighbours have pets?

Mary: Some of them do. (Sperber & Wilson, 1995, p. 277)

The interpretation *some (and possibly all) of Mary's neighbours have pets* is relevant enough, but not the most relevant one, so if this interpretation is reached, processing should continue. The interpretation should then be reached that some but not all of Mary's neighbours have pets.

Assuming that it is right that there is a presumption of optimal relevance, then there is a sense in which 'blind' accessibility-only accounts are accidentally close to being correct. As discussed, the presumption of optimal relevance mandates following a least-effort path, and the first interpretation that comes to mind must lie on that path. So the most accessible solution will at least be on the path to the solution, and, given that the attunement of our memory structure to the world is part of the reason for the frugality of comprehension, the most accessible interpretation will often be the intended solution. There will however be cases in which the most accessible interpretation is not correct and should be rejected in favour of an interpretation that, while on the

without reference to Lavie's perceptual-load theory) that the correct account must be a hybrid between 'early selection' and 'late selection' theories.

[211] Such attentional factors may be top-down or bottom-up. Attention can be consciously directed, but can also be captured by highly salient stimuli.

least-effort path, is a bit harder to reach. This will occur when either the most accessible solution is not relevant enough even for the small effort required in reaching it (by clause a) or there is reason to suppose that putting a bit more effort in will bring significantly higher returns (by clause b), or both.

The second way that frugality is achieved by constraints on the solutions generated is related to a second problem with a non-inferential account. In a non-inferential account there is no licence to generate only solutions where the explicit meaning, together with implicated premises, logically supports the implicated conclusions. A theory like Recanati's is committed to the claim that accessibility and salience will deliver the correct explicit meaning of utterances without this constraint on the trial interpretations generated.

For Recanati the two types of process involved in pragmatics, primary and secondary, are analogous to visual processing of a scene, followed by inferences about what is perceived. (For example, Holmes perceives a rope hanging down from the ceiling to the bed, and infers that the snake came down it.) While the two processes may happen close to simultaneously, the former feeds the latter, and the former is a pseudo-inferential 'brute' process, while the latter is a properly inferential process. The coherence of one's inferences with what one perceives is a criterion for accepting or rejecting those inferences, but it is not a criterion for rejecting one's perceptions, barring certain exceptional circumstances like optical illusions. Such illusions in fact demonstrate that one cannot generally change what one perceives on the basis of reflective inference, even when the inference suggests that one should. Another way of putting this point is that there are certain constraints on the solutions generated by perceptual processing, and other constraints on inference, but these constraints (except ones that apply to cognition in general) are not shared or met jointly.

In contrast, there is evidence suggesting that the implicatures and the explicit meaning of an utterance must be generated in tandem, because there are constraints on the interpretation of an utterance as a whole. There are examples that show that explicit content is enriched to just the degree required to support a particular implicature.

(23) Peter: Do you want to go to the cinema?

Mary: I'm tired. (The example is from Sperber & Wilson, 1998. See also Recanati, 2004, p. 47 for discussion.)

Here the explicit meaning conveyed by Mary's utterance is that she is tired at least to a degree that makes her not want to go to the cinema. This logically supports an implicature to the effect that she does not want to go to the cinema, which answers the question Peter posed. Given the contextual expectation that the question raises, without that implicature the utterance would not be sufficiently relevant, and without the enrichment of the explicit meaning, the implicature would not be warranted.

This kind of interdependence between explicit meaning and implicatures suggests that there must be mutual adjustment during their derivation. That in turn suggests that if the derivation of one is inferential, then the derivation of both must be. Recanati accepts mutual adjustment, but denies that it is a problem for his theory (Recanati, 2004, pp. 46–50). However, he does not explain how the content of implicatures in his picture could affect the content of the explicit meaning. Perhaps one could argue that the facts do not show that the derivation of explicit meaning is inferential, and that they show only that primary pragmatic processing (unlike visual processing) is not encapsulated. Then top-down suppression or activation from the result of secondary processing might influence the blind primary processing. This argument, in my opinion, remains to be made.

I am more interested, here, in the strong constraint on interpretations that this kind of mutual adjustment suggests. All interpretations must be such that the explicit meaning – together, perhaps with implicated premises – warrants (i.e. logically supports) the implicatures or implications that make the utterance relevant in the expected way. This, in its way, is as strong a constraint as the constraint discussed in chapter 3 that a maximum must lie on a particular curve. The analogy is rather precise. In both cases, the existence of the constraint means that it is rational to generate only solutions which satisfy the constraint. Further, the form of the constraint suggests how such solutions should be generated. If a solution lies on a curve, then the equation of the curve can be used to generate trial solutions. Similarly, given that explicatures must warrant implicatures or implications, that is, that there must be a logical argument with explicit meaning as premise and the implicatures or implications as conclusions, the generator for such conclusions can be simple forward inference from the explicit meaning. I have described this kind of process in chapter 3.

There is one further point of similarity. The starting point for the generation of solutions in both cases is *not* determined by the constraint. There must be some choice made of where on the curve to start looking for the maximum, and there must be some choice made of what enrichment of the encoded meaning to start with, and what other premises to attempt to combine with the explicit meaning to derive an implicated conclusion. In the generation of potential interpretations, the system can rely on the other very strong constraint already described, that a least-effort path is followed. That constraint mandates the use of the initially most accessible supplementary premises and enrichment of the explicit meaning as a starting point: on a least-effort path, you start where you already are.

The discussion above can be taken as my answer to an objection to the view that input to utterance interpretation is propositional. Opponents of this view might regard it as a kind of sleight of hand to say that the result of parsing is embedded under something like "S said that..." and that this is the kind of input taken by pragmatic processing. They might say that one should be able to do the same for parsing or visual processing, or any pseudo-inferential process, and then claim that that process is inferential because it works with conceptual input and output. My reply is that one could indeed postulate that the input to any of these processes is embedded into a conceptual representation in a similar way, but that there is no reason to make this postulation unless doing so would be fruitful. I have tried to show how it is fruitful in theorising about pragmatic processing to assume that the linguistic input is embedded in a conceptual representation: it allows the postulation of certain strong constraints that are supported by evidence of mutual adjustment between implicatures and explicatures. Theorists of linguistic parsing or visual inference could make a similar move if it seemed likely to be productive to do so[212]. I suspect that they do not (Helmholtz, perhaps, aside – see §4.3.5 above) because it is clear to them that it would not be.

One might suggest that this is an *a priori* issue: there should be some consideration that says that (e.g.) visual processing or linguistic parsing could not be inferential. One of the points I have tried to make in the previous chapter is that I do not think the introspective or 'personal' status of a process are considerations that will do this job. It is more relevant that utterances are actions and may provide clues to relevant information (in the first instance, about the speaker's intentions, as discussed), while sentences do not do so in their own right, but only as uttered by speakers. The postulate that linguistic information is embedded into conceptual representation reflects this point.

5.3 HEURISTICS, MODULES AND EFFICIENCY

Rationality implies efficiency, as Simon said (see the epithet to this chapter). Utterance interpretation (and most utterance production) takes place under such severe time pressure that the efficiency of the processes involved becomes a central question. The discussion here of the constraints on the generation of trial interpretations is intended to show that it is plausible that a computational process can be efficient enough.

In general, the claim that a heuristic proposed for an inference process is frugal implies a claim that Fodor's problem is not fatal for that kind of inference. Heuristics can ignore information, where an algorithmic procedure would have to take all information into account (and ideal visions of rationality simply assume that conclusions are reached as though all information were somehow considered). Heuristics that are shortcuts, for a certain type of problem, in a certain domain – that is, adaptive heuristics – ignore information in such a way that they are faster than algorithmic alternatives and more

[212] Although doing so would neglect an important difference between perception and comprehension. It is utterances (or facts about them) that are found to be relevant, not speech sounds or sentences (or even sentence tokens). Phonetic representations have little intrinsic relevance to the hearer, unlike light impinging on his retina or sounds impinging on his ears. My thanks to Deirdre Wilson (p.c.) for pointing out the relevance of this point here. See also the next paragraph in the text.

parsimonious with resources like representation and processing, while delivering answers that are accurate enough, enough of the time.

Recall that Fodor's contention is that essentially no central (i.e. conceptual) processing involving non-demonstrative inference can be explained computationally, because a) any information might be relevant, since it might be evidence bearing on the conclusion, and b) the consequences of any postulated conclusion should in principle be evaluated in terms of their effect on the global properties of the inferer's belief system, such as its overall coherence. His worries do not extend to peripheral processes such as visual processing because these are informationally encapsulated. We can know, for example, that the two horizontal lines in the Müller-Lyer diagram are the same length, without this altering our perception that one of them is longer. Information from long-term memory and from reflective inference cannot (in general) affect the results delivered by peripheral systems.

In recent years many theorists have suggested that central, conceptual processing has some structures of a modular or quasi-modular kind, and that aspects of central cognition can thus be separated out and studied. I do not want to enter the massive modularity wars in this thesis. My aim in this section is to show that fast, automatic processes can solve some of the problems that modules are supposed to solve. The fast and frugal heuristics programme and work on central modularity, are, in my opinion two research programmes aimed at the same problem. Whether one describes an adaptive heuristic for a domain as a module depends on one's definition of the term 'module'.

Fodor modules are strongly encapsulated processes, or bundles of processes. A different type of strict modularity is possessed by Chomsky modules (this name for them is from Segal, 1996). A Chomsky module is a domain-specific database, i.e. a body of knowledge specific to a domain, as with knowledge of language. This second kind of strict modularity may be possessed by some central modules. Tsimpli and Smith (1998, p. 212) argue that the faculty of language is partly a central module in this sense.[213] If a certain kind of pro-

[213] This relies on a different view of centrality from mine. The consideration here is that the same database is involved in parsing and constructing sentences, input and output processes: no one speaks only English and understands only Icelandic. In this thesis, centrality is defined by the type of input taken by a process. If a process takes conceptual input it is central, by definition.

cessing only consults a domain-specific database, then it has a better chance of avoiding computational explosion.

I mention Fodor and Chomsky modules as solutions to the problem of computational explosion in order to set them aside. Pragmatic processing is not informationally encapsulated in Fodor's sense: any information may be relevant[214]. The evidence is that general knowledge is used in pragmatic processing: e.g. in (21c) above, the referent of 'he' may be determined partly by knowledge of the degree to which the local police are corrupt. In addition, there does not seem to be mentally represented knowledge specific to pragmatics: the constraints and regularities I have described are properties of the process, not knowledge that is used to guide processing.[215]

I claim that for conceptual processes such as utterance interpretation there is a different type of solution to Fodor's problem. Informational encapsulation is the key property of a Fodor module, but frugality and access to information are orthogonal issues.[216] All four logical possibilities are exemplified by some kind of processing. There are frugal, encapsulated processes. Visual processing is quick and mostly encapsulated, as is linguistic parsing. The ball-catching heuristic discussed in chapter 1 is very frugal and mostly encapsulated. The heuristic appears only to represent one quantity, and works that quantity out from a simple observation. One's procedure for catching a ball is not sensitive to general inference. (The speed with which the process occurs may be largely responsible for this). Spontaneous deductions in reasoning problems are frugal (assuming that generation of trivial implicatures is ruled out), and in high-IQ participants they are (deliberately or habitually) encapsulated against such considerations as the plausibility of the conclusion.

[214] Indeed in a central architecture where modules are interlinked it is not clear what informational encapsulation would amount to. Informational encapsulation is really the condition that information cannot affect peripheral processing from above: cross-modal phenomena such as the Stroop effect are not in contradiction with it, but effects on perception from beliefs reached by inference would be.

[215] See Kasher's work however for an attempt to deal with part of pragmatics as a Chomskyan competence – a domain-specific body of knowledge, e.g. Kasher, 1991.

[216] Since the question of whether the faculty for processing in a particular domain is modular is orthogonal to the question of whether it avoids computational explosion, it is not surprising that Sperber and Wilson have changed their position on the modularity of pragmatic processing without changing their position that pragmatic processing is not prone to Fodor's problem.

There are also processes that use no external information but require huge effort. Truth-table calculations only take into account the input information, but they are far from frugal. Playing chess uses heuristics that can mostly only take into account the state of the board and knowledge of previous games, but which are not usually quick.

There are of course also areas of life where cognition is non-encapsulated and non-frugal. Scientific investigation, Fodor's favourite example of central cognition, is an obvious case in point, although even here much of the thought involved may be fast, heuristic and frugal. Fodor's argument rests on an analogy between what scientists do (theory choice) and what individuals do in everyday reasoning, which has recently been challenged by Carruthers (2003) and Pinker (2005) as well as consistently by Sperber and Wilson (1986; 1996).[217]

It is the fact that no theory is permanently established that makes it possible for science eventually to take into account any and all information. As Sperber and Wilson say (1986, p. 166), given that the aim is the best possible theory, there is no criterion by which a hypothesis can be permanently established except comparison with all competitors, so exhaustive search is necessary, and the domain of possible solutions is open-ended, so search continues indefinitely.

Finally there is the class of processes which I am most interested in, those which are not encapsulated but still frugal. The 'take the best' heuristic is an example. It is not encapsulated in any strong sense. It can consult any cue in deciding which is the better alternative, and if the first cue it looks at is not decisive it looks at another, and so on. Thus on some occasions it will consult all of the available cues before making a decision. (In a real-world situation, if no decisive cue were found, more information would be sought, or decision deferred.) On average, however, it consults very few cues, so it is more frugal than the classical alternative, a weighted calculation across all the cues provided or available.

My point is that adaptive heuristics are not prone to Fodor's problem because they systematically ignore certain information, but not necessarily by

[217] I discussed in chapter three some of the points Sperber and Wilson make. See also the main text below.

being encapsulated: rather, they do it by taking a limited amount of information in a particular order and stopping soon. The relevance-theoretic comprehension procedure can consult – in principle – any information, in its task, inferring the best explanations of ostensive stimuli. This means that by definition, the comprehension procedure is a quasi-module in Smith and Tsimpli's sense. It is domain specific, but not informationally encapsulated (Smith & Tsimpli, 1995; Tsimpli & Smith, 1998). Its frugality, however, does not follow from its quasi-modularity. It should be frugal because of the strong constraints and recognition factors discussed above.

5.3.1 PROCESS ENCAPSULATION

Notwithstanding the existence of this dedicated (quasi-)module, it is possible to bring other processes to bear on utterance interpretation. As discussed above, if normal, fast processing fails to produce a result, slow, reflective processing may be employed. Such processing will have similarities to fast processing, but certain presumptions may be suspended: perhaps what the hearer is intended to find relevant is the fact that there are two potential interpretation, as in puns; perhaps the speaker was so confused or such a bore that the utterance falls massively short of relevance and the least effort path will not lead to the intended interpretation.[218] Another possibility is that the piece of behaviour or sound taken as an utterance was not intentional so there is no intended interpretation: in such cases the best explanation lies elsewhere.

My suggestion is that there is a kind of process encapsulation. Input in a certain domain activates and is then processed by a certain heuristic or set of heuristics, fast and nearly automatically. On short timescales the process applied to a particular kind of problem or stimulus is inflexible. With a bit more time, and, possibly, conscious attention, the procedure might be varied to a greater or lesser degree. This proposal is a way of accounting for the fact that

[218] In principle the relevance-theoretic comprehension procedure and the fact that the hearer's expectations of relevance may be revised in the course of the interpretation process allows for such utterances. At the start, the interpretation is presumed to be the most relevant one the speaker was willing and able to convey; but a speaker may only intend an utterance to seem relevant, and interpretation may be successful nonetheless. However, it is intuitively clear that a certain level of conversational incompetence or selfishness sometimes defeats fast comprehension by the normal route.

much rapid thought appears to follow similar tracks, pursuing specific aims, whereas at least some conscious, effortful thought has a more open-ended character.

In a paper defending Fodor's analogy between central cognition in general and scientific investigation, Dominic Murphy claims that advocates of central modularity "bet that the world is divided up into domains that do not constrain each other, and that our mind mirrors that structure" (2006, p. 564). I suggest that the bet that one should make is rather different: that for some decisions, in some domains, at some time-scales, the mind behaves in a way that would work well if the world is compartmentalised in the way suggested. The world may be like that or not: the claim is about psychology, not about the ontology of non-mental parts of the world.

My proposal can be construed as the claim that modularity of central processes is relative to timescale.[219] We can integrate across domains or bring a fresh approach to a problem in a domain, but generally only if we have plenty of time. I think that there is evidence for this view from pragmatic illusions.

Perceptual and cognitive illusions

I have mentioned optical illusions. What happens in a perceptual illusion (there are illusions for other sense modalities as well as vision) is that information held in the mind, even in active, short-term memory, cannot influence the result of processing by a perceptual module. There are phenomena that are in some ways similar in utterance interpretation, known in the psycholinguistics literature as semantic illusions (Erickson & Mattson, 1981; van Oostendorp & De Mul, 1990; van Oostendorp & Kok, 1990; Barton & Sanford, 1993). For example, in the middle of a questionnaire, participants are asked "How many animals of each kind did Moses take into the ark?" The majority give an answer, rather than noting that the question is odd, since it is Noah, not Moses, who is associated with the ark. Even when participants are told about such illusions and asked to look out for them, they still fall victim, al-

[219] This would make modularity similar to the property of adiabacity. In thermodynamics, an adiabatic process is one in which no heat is transferred to or from the system under consideration. The term "adiabatic" literally indicates an absence of heat transfer, but a transformation of a thermodynamic system is considered adiabatic when it is quick enough so that no significant heat transfer happens between the system and the outside.

though at a lower rate. Allott and Rubio Fernández (2002) argued that illusions of this type are pragmatic rather than semantic illusions, created by shallow processing in utterance interpretation. As with other pragmatic phenomena, the context should play a crucial role in the interpretations reached, and this role should not be explicable purely in terms of stereotypical knowledge such as frames or schemas. The experiments we suggested have not been carried out, but I assume here that it is correct that these illusions are essentially pragmatic and conceptual in nature.

Illusions caused by shallow processing in pragmatics, while related to perceptual illusions, are significantly different from them. They differ in that the problem is not that the pragmatic faculty does not have potential access to the right information but that it reaches a decision without considering some highly relevant information that it could have considered. The fatal ignition/illumination mistake discussed in chapter 3 can be analysed similarly as due to shallow processing – in that case by a central process concerned with action plans. What I am suggesting is that this difference between these two kinds of illusion is due to the difference between the underlying mental apparatus: for perceptual processing, a Fodor module which cannot consult certain sources of information; for pragmatic processing a fast process that can in principle consult any information, but in practice has rules which rapidly curtail search.

Perceptual illusions such as the Müller-Lyer illusion are persistent. The perceptual feeling that the top line is longer than the other line does not go away after measurement. Pragmatic illusions are not like this. Once you have properly understood that the word was 'Moses' the question no longer feels as though it is asking about Noah. So one's other beliefs can overturn pragmatic illusions. However they are persistent in another way. Even when one is told to look out for such illusions it is hard to avoid falling into the trap.[220]

I would like to suggest that these phenomena demonstrate what is effectively a 'soft' form of informational encapsulation, contrasting with proper,

[220] This may be hard to believe, but the experimental results demonstrate it, as does anecdotal evidence. I have carefully explained pragmatic illusions to a professor of linguistics, immediately asked the 'Moses' question, received the illusory answer, explained what is wrong with it, then asked a second pragmatic illusion question, again getting the illusory answer. I stopped at that stage, but feel that it might have been possible to continue.

'hard' encapsulation.[221] In hard encapsulation, the module keeps doing its job the same way, delivering the same result regardless of the agent's central or conscious beliefs, so one line in the Müller-Lyer diagram, for example, always seems shorter than the other. Soft encapsulation occurs because a central process is fast and frugal. Such a process aims to deliver a result quickly. If a result is found that is relevant in the expected way, then processing is ended, in some cases prematurely.

On my view, in pragmatic processing the information considered in processing any input is stipulated by the process in a way that cannot be changed very much during processing. This latter property is not full-blown informational encapsulation, but it is clearly something quite different from Fodor's view of central processes performing holistic abductive inference and decision-making. In particular it is quite different from scientific theory choice.

It has been notoriously hard to say what the scientific method is, as the history of the philosophy of science testifies. One thing that is clear is that conscious reflection on the problems faced at a particular time can modify the methods used by scientists. In this sense there is no heuristic of scientific method, although there are no doubt rules of thumb within particular fields of science (e.g. presented with an unknown chemical, look at the colour of flame it produces when heated).

One difference between ordinary central cognition and full-blown scientific reasoning, on this thesis, is that while both involve abductive inference, ordinary central cognition is largely a matter of following heuristics at timescales fast enough that the results of the processing do not significantly change the heuristic, whereas in science a significant role is played by changes in the methods of investigation used, driven by reflection on the results that are being produced.

[221] Carruthers (2007) makes a related distinction between narrow-scope and wide-scope encapsulation. Another useful comparison is with recent work by Hauser (2003) on modular macros: "fast, automatic, unconscious action sequences" (Hauser, 2003, p. 80).

5.4 CONCLUSION

In this final chapter I have argued (following Sperber and Wilson 1986; 1996) that the pragmatic system works out speaker's meaning frugally, zeroing in rapidly on the relevant interpretation and the assumptions needed to support it. It is relatively fast and frugal, following a least-effort path, and it has a rather singular dynamic aspiration level, exploiting the task environment of inferential-intentional comprehension. The process makes use of the way that the problem constrains possible solutions. It generates, I suggest, only hypothesised solutions that stand in the right logical relationship to the facts to be explained. Crucially, it is a system that does not fall victim to Fodor's problem.

Fodor's claim, as I have explained, is that central cognition must somehow take into account all potentially relevant information because that is the only way to be sure that one has found the best solution. It has been said that the only cognitive systems that do not fall foul of Fodor's problem are modular systems. I have argued (again agreeing with Sperber and Wilson) that the questions of modularity and computational explosion are orthogonal to each other. Strong modularity is the property of cognitive systems that are informationally-encapsulated and domain-specific. I have claimed that one reason that utterance interpretation is not subject to computational explosion is that it is subject to what might be called process encapsulation. The process that is followed in the initial attempt to work out the interpretation of an utterance, while inferential, is reflex-like in that it is fast and that major elements are uniform across utterances.

I have not outlined a comparable utterance construction procedure in this chapter. As I commented in chapter 4, an utterance is a kind of action, and like other actions is directed towards fulfilment of a goal or end. As Adam Morton writes:

> Normally there are infinitely many means to one's end, the best of which one has not thought of. So, almost always, there is nothing that is necessary for one's end, and a confusing set of things that are sufficient. What rationality demands is that one go through the few that one can focus one's mind on and find an acceptable one. (Morton, 2006, p. 771)

More work on production is required, but given that production operates under similar constraints to interpretation, I think that the underlying processes are parallel to interpretation at least in that a) working out what utterance will convey roughly the desired meaning is a kind of reasoning; b) it is accomplished by aspiration-level terminated sequential search; c) often, but not always, the first acceptable solution generated will be the one chosen.

The aim in this chapter, and in the thesis as a whole, has been to show how realistic views of rationality and a broadly Gricean inferential-intentional view of pragmatics can cohere, and what light they shed on each other. I hope to have made the case that a bounded view of rationality is necessary, and that this view is compatible with the assumption that communication is inferential. I have shown how I believe that these views lead to a view of utterance interpretation as carried out by a certain kind of heuristic search procedure.

APPENDIX I

Grice's 'expansion' of Shropshire's 'argument'

If the soul is not dependent on the body, it is immortal.

If the soul is dependent on the body, it is dependent on that part of the body in which it is located.

If the soul is located in the body, it is located in the head.

If the chicken's soul were located in its head, the chicken's soul would be destroyed if the head were rendered inoperative by removal from the body.

The chicken runs round the yard after head-removal.

It could do this only if animated, and controlled by its soul.

So the chicken's soul is not located in, and not dependent on, the chicken's head.

So the chicken's soul is not dependent on the chicken's body.

So the chicken's soul is immortal.

If the chicken's soul is immortal, *a fortiori* the human soul is immortal.

So the soul is immortal.

(Grice, 2001, pp. 11–12)

APPENDIX II

Time taken to check consistency of truth-table for 100 independent propositions

Number of rows = 2^n : n=100

$= 2^{100} = 1.27 \times 10^{30}$

Assume that rate of checking = 10 rows per second

Number of seconds in a year = 60 × 60 × 24 × 365 = 31,536,000

Years to check all rows = (number of rows)/(rate of checking × seconds in a year)

$= 1.27 \times 10^{30}/(10 \times 31{,}536{,}000)$

$= 4.03 \times 10^{21}$

REFERENCES

Allott, N. (2005). Paul Grice, reasoning and pragmatics. *UCL Working Papers in Linguistics, 17*, 217–243.

Allott, N. (2006). Game theory and communication. In A. Benz, G. Jäger & R. van Rooij (Eds.), *Game Theory and Pragmatics.* (pp. 123–151). Basingstoke: Palgrave Macmillan.

Allott, N. & Rubio Fernández, P. (2002). This paper fills a much-needed gap. *Actes de l'atelier des doctorants en linguistique, Université Paris 7*, 97–102.

Antoniou, G. & Williams, M. A. (1997). *Nonmonotonic Reasoning.* Cambridge, Mass: MIT Press.

Antony, L. M. & Hornstein, N. (2003). Introduction. In L. M. Antony & N. Hornstein (Eds.), *Chomsky and his Critics.* (pp. 1–10). Malden, MA: Blackwell.

Aristotle. (1998). *The Nicomachean Ethics* (W. D. Ross, J. L. Ackrill & J. O. Urmson, Trans.). Oxford: Oxford University Press.

Audi, R. (2001). *The Architecture of Reason: The Structure and Substance of Rationality.* New York: Oxford University Press.

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology, 36*(2), 147–168.

Barlow, H. (2002). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences, 24*(4), 602–607.

Barnard, C. I. (1968). *The Functions of the Executive* (30th anniversary ed.). Cambridge, Mass: Harvard University Press. (Originally published 1938.)

Baron-Cohen, S., Leslie, A. M. & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.

Barton, E. L. (1990). *Nonsentential Constituents: A Theory of Grammatical Structure and Pragmatic Interpretation.* Amsterdam: John Benjamins.

Barton, S. B. & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition, 21*(4), 477-487.

Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences, 16*(2), 337–338.

Båve, A. (2008). A pragmatic defense of Millianism. *Philosophical Studies, 138(2)*, 271–289.

Behne, T., Carpenter, M., Call, J. & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*(2), 328–337.

Benz, A., Jäger, G., & van Rooij, R. (Eds.). (2006). *Game Theory and Pragmatics.* Basingstoke: Palgrave Macmillan.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press.

Bloom, P. & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25–B31.

Bonini, N., Tentori, K. & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language, 19*(2), 199–210.

Boole, G. (1854). *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities.* London: Walton and Maberly.

Borges, B., Goldstein, D. G., Ortmann, A. & Gigerenzer, G. (1999). Can ignorance beat the stock market? In G. Gigerenzer, P. M. Todd & ABC Research Group (Eds.), *Simple heuristics that make us smart.* (pp. 59–72). New York: Oxford University Press.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*(1), 1–21.

Braine, M. D. S. (1990). The "natural logic" approach to reasoning. In W. F. Overton (Ed.), *Reasoning, Necessity, and Logic: Developmental Perspectives.* (pp. 133–157). Hillsdale, NJ: Erlbaum.

Braine, M. D. S. & O'Brien, D. P. (Eds.). (1998). *Mental Logic.* Mahwah, N.J: L. Erlbaum Associates.

Braine, M. D. S., Reiser, B. J. & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. *The Psychology of Learning and Motivation, 18*, 313–371.

Brancazio, P. J. (1985). Looking into Chapman's homer: The physics of judging a fly ball. *American Journal of Physics, 53*, 849.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review, 93*(3), 375–405.

Bratman, M. (1987). *Intention, Plans and Practical Reason.* Cambridge, Massachusetts: Harvard University Press.

Breheny, R. (2003). On the dynamic turn in the study of meaning and interpretation. In J. Peregrin (Ed.), *Meaning in the Dynamic Turn.* (pp. 69–90). Dordrecht: Elsevier.

Breheny, R. (2006). Communication and folk psychology. *Mind & Language, 21*(1), 74–107.

Brown, H. I. (1988). *Rationality.* London: Routledge.

Burge, T. (1993). Content preservation. *The Philosophical Review, 102*(4), 457–488.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*(1), 61–83.

Byrne, R. M. J. (1991). Can valid inferences be suppressed? *Cognition, 39*(1), 71–78.

Byrne, R. M. J., Espino, O. & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*(3), 347–373.

Byrne, R. M. J., Espino, O. & Santamaría, C. (2000). Counterexample availability. In W. Schaeken, G. De Vooght & G. d'Ydewalle (Eds.), *Deductive Reasoning and Strategies.* (pp. 97–119). London: Lawrence Erlbaum.

Campbell, R. (1981). Language acquisition, psychological dualism and the definition of pragmatics. In H. Parret, M. Sbisà & J. Verschueren (Eds.), *Possibilities and Limitations of Pragmatics.* (pp. 93–103). Amsterdam: John Benjamins B.V.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G. & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*(4), i–174.

Carruthers, P. (2003). On Fodor's problem. *Mind & Language, 18*(5), 502–523.

Carruthers, P. (2007). Simple heuristics meet massive modularity. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The Innate Mind: Culture and Cognition.* (pp. 181–198). Oxford: Oxford University Press.

Carston, R. (1988). Implicature, explicature and truth-theoretic semantics. In R. Kempson (Ed.), *Mental Representations: The Interface Between Language and Reality.* (pp. 155–181). Cambridge: Cambridge University Press.

Carston, R. (2000). The relationship between generative grammar and (relevance-theoretic) pragmatics. *Language and Communication, 20,* 87–103.

Carston, R. (2002a). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language, 17*(1/2), 127–148.

Carston, R. (2002b). *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell.

Carston, R. (2003). Conversational implicatures and pragmatic mechanisms, abstract for the conference of the European Society for Philosophy and Psychology, 2003. 2005-08-02 edition. Retrieved 2005-08-02, http:/ /www.eurospp.org/2003/papers/Doc2003/Carston206%20LIN%20.doc

Cheng, P. W. & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391–416.

Cheng, P. W. & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition, 33*(3), 285–313.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E. & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18*(3), 293–328.

Cherniak, C. (1981). Minimal rationality. *Mind, 90*(358), 161–183.

Cherniak, C. (1986). *Minimal Rationality.* Cambridge, Mass: MIT Press.

Chomsky, N. (1959). A review of B. F. Skinner, 'Verbal Behavior', 1957. *Language, 35,* 26–58.

Chomsky, N. (1964). *Current Issues in Linguistic Theory.* The Hague: Mouton.

Chomsky, N. (1966). *Cartesian Linguistics.* Harper & Row.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use.* New York: Praeger.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures.* Cambridge, Mass: MIT Press.

Chomsky, N. (1991a). Linguistics and adjacent fields: A personal view. In A. Kasher (Ed.), *The Chomskyan Turn.* (pp. 3–25). Oxford: Basil Blackwell.

Chomsky, N. (1991b). Linguistics and cognitive science: Problems and mysteries.

In A. Kasher (Ed.), *The Chomskyan Turn.* (pp. 26–55). Oxford: Basil Blackwell.

Chomsky, N. (1992). Language and interpretation: Philosophical reflections and empirical enquiry. In J. Earman (Ed.), *Inference, Explanation and Other Frustrations: Essays in the Philosophy of Science.* (pp. 99–128). Berkeley, CA: University of California Press.

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1–61.

Chomsky, N. (2003). Replies. In L. M. Antony & N. Hornstein (Eds.), *Chomsky and his Critics.* (pp. 253–328). Malden, MA: Blackwell.

Clark, H. (1977). Bridging. In P. Johnson Laird & J. Wason (Eds.), *Thinking: Readings in Cognitive Science.* (pp. 411–420). Cambridge University Press.

Clements, W. A. & Perner, J. (1984). Implicit understanding of belief. *Cognitive Development, 9,* 377–395.

Cohen, L. J. (1981). Can human rationality be demonstrated experimentally? *Behavioral and Brain Sciences, 4,* 317–370.

Cohen, L. J. (1992). Rationality. In J. Dancy & E. Sosa (Eds.), *A Companion to Epistemology.* (pp. 417–420). Cambridge, Mass: Blackwell.

Collins, P. D. B., Martin, A. D. & Squires, E. J. (1989). *Particle Physics and Cosmology.* New York: Wiley.

Conan Doyle, A., Sir. (1892a). Silver Blaze (from 'The Memoirs of Sherlock Holmes'). Retrieved 10-12-06, http://en.wikisource.org/wiki/Silver_Blaze

Conan Doyle, A., Sir. (1892b). The Adventure of the Speckled Band (from 'The Adventures of Sherlock Holmes'). Retrieved 16-12-06, http:/ /en.wikisource.org/wiki/The_Adventure_of_the_Speckled_Band

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature, 34*(2), 669–700.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187–276.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* (pp. 163–228). Oxford: Oxford University Press.

Cosmides, L., Tooby, J., Fiddick, L. & Bryant, G. A. (2005). Detecting cheaters. *Trends in Cognitive Science, 9*(11), 505–6; author reply 508–10.

Dancy, J. (2003). Aspects of reason I (Review of 'Aspects of Reason', Paul Grice, 2001). *The Philosophical Quarterly, 53*(211), 274–279(6).

Davidson, D. (1963). Actions, reasons and causes. *Journal of Philosophy, 60,* 685–700.

Davidson, D. (1980a). *Essays on Actions and Events.* Oxford: Clarendon Press.

Davidson, D. (1980b). Psychology as philosophy. In *Essays on Actions and Events.* (pp. 229–244). Oxford: Clarendon Press.

Day, J. M. (Ed.). (1994). *Plato's Meno in Focus.* London: Routledge.

Dehaene, S. (1999). *The Number Sense: How the Mind Creates Mathematics.* London: Penguin. (Originally published Oxford: Oxford University Press, 1997.)

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.

Dennett, D. C. (1987). *The Intentional Stance.* Cambridge, Mass: MIT Press.

Descartes, R. (1912). *A Discourse on Method Etc.* (J. Veitch, Trans.). London: Everyman's Library; J M Dent. (Originally published 1637.)

Dienes, Z. & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*(05), 735–808.

Douven, I. & Uffink, J. (2003). The preface paradox revisited. *Erkenntnis, 59*(3), 389–420.

Dreier, J. (2004). Decision theory and morality. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 156–181). New York: Oxford University Press.

Dulany, D. E. & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition, 9*(1), 85–110.

Dummett, M. A. E. (1973). *Frege: Philosophy of Language.* London: Duckworth.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics, 75*(4), 643–669.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality.* Cambridge: Cambridge University Press.

Enfield, N. J. & Levinson, S. C. (2006). Introduction: Human sociality as a new interdisciplinary field. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction.* (pp. 1–35). Oxford: Berg.

Erickson, T. D. & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 540-551.

Evans, J. S. B. T., Handley, S. J., Harper, C. N. J. & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(6), 1495–1513.

Evans, J. S. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology, 24,* 193–199.

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75*(4), 541–568.

Evans, J. S. B. T. (1989). *Bias in Human Reasoning: Causes and Consequences.* London: Erlbaum.

Evans, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*(2), 223–240.

Evans, J. S. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning, 4*(1), 45–110.

Evans, J. S. B. T. (1999). The influence of linguistic form on reasoning: The case of matching bias. *The Quarterly Journal of Experimental Psychology: Section A, 52*(1), 185–216.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science, 7*(10), 454–459.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378–395.

Evans, J. S. B. T., Barston, J. L. & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*(3), 295–306.

Evans, J. S. B. T., Ellis, C. E. & Newstead, S. E. (1996). On the mental representation of conditional sentences. *The Quarterly Journal of Experimental Psychology: Section A, 49*(4), 1086–1114.

Evans, J. S. B. T. & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*(3), 391–397.

Evans, J. S. B. T., Newstead, S. E. & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction.* Hove: Psychology Press.

Evans, J. S. B. T. & Over, D. E. (1996). *Rationality and Reasoning.* Hove: Psychology Press.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*(2), 123–129.

Fodor, J. A. (1975). *The Language of Thought.* New York: Crowell.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology.* Cambridge, Mass: MIT Press.

Fodor, J. A. (1985a). Précis of the modularity of mind. *Behavioral and Brain Sciences, 8*(1), 1–42.

Fodor, J. A. (1985b). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind, 94*(373), 76–100.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, Mass: MIT Press.

Fodor, J. A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology.* Cambridge, Mass: MIT Press.

Fodor, J. A. (2005). Reply to Steven Pinker 'So how does the mind work?'. *Mind & Language, 20*(1), 25–32.

Fodor, J. A. (2007). The revenge of the given. In B. McLaughlin & J. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind.* Blackwell.

Frege, G. (1964). *The Basic Laws of Arithmetic; Exposition of the System* (M. Furth, Trans.). Berkeley: University of California Press.

Frege, G. (1979). *Posthumous Writings* (P. Long & R. White, Trans.). Chicago: University of Chicago Press.

Frege, G. (1984). *Collected Papers on Mathematics, Logic, and Philosophy* (M. Black, Trans.). Oxford: B. Blackwell.

Gamut, L. T. F. (1990). *Logic, Language and Meaning: Volume 1, Introduction to Logic.* Chicago: University of Chicago Press.

Garcia-Carpintero, M. (2001). Gricean rational reconstructions and the semantics-pragmatics distinction. *Synthèse, 128*, 93–131.

Garner, B. A. & Black, H. C. (2004). *Black's Law Dictionary* (8th ed.). St. Paul, MN: Thomson/West.

Gauthier, D. P. (1986). *Morals By Agreement.* Oxford: Clarendon Press.

Gentzen, G. (1964). Investigations into logical deduction. *American*

Philosophical Quarterly, 288–306.

Gergely, G., Bekkering, H. & Király, I. (2002). Rational imitation in preverbal infants. *Nature, 415*(6873), 755.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*(3), 592–596.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World.* New York: Oxford University Press.

Gigerenzer, G. (2001). Decision making: Non-rational theories. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences, Volume V.* (1st ed., pp. 3304–3309). Amsterdam: Elsevier.

Gigerenzer, G. (2004). Striking a blow for sanity in theories of rationality. In M. Augier & J. G. March (Eds.), *Models of a Man: Essays in Memory of Herbert A. Simon.* (pp. 389–409). Cambridge, Mass: MIT Press.

Gigerenzer, G. (2005). I think, therefore I err. *Social Research: An International Quarterly of Social Sciences, 72*(1), 1–24.

Gigerenzer, G., Czerlinski, J. & Martignon, L. (2002). How good are fast and frugal heuristics? In R. Elio (Ed.), *Common sense, reasoning, & rationality.* (pp. 148–173). New York: Oxford University Press.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*(4), 650–669.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506–528.

Gigerenzer, G. & Selten, R. (Eds.). (2001). *Bounded Rationality: The Adaptive Toolbox.* Cambridge, Mass: MIT Press.

Gigerenzer, G. & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart.* New York: Oxford University Press.

Gilhooly, K. J. (2004). Working memory and reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 49–77). Cambridge: Cambridge University Press.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E. & Wynn, V. (1993). Working memory and strategies in syllogistic-reasoning tasks. *Memory and Cognition, 21*(1), 115–124.

Girotto, V., Kemmelmeier, M., Sperber, D. & van der Henst, J. B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition S, 81*(2), B69–76.

Goldman, A. I. (1970). *A Theory of Human Action.* Englewood Cliffs, N.J: Prentice-Hall.

Goldstein, D. G. & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd & the ABC Research Group (Eds.), *Simple Heuristics That Make Us Smart.* (pp. 37–58). New York:

Oxford University Press.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75–90.

Goodman, N. (1954). *Fact, Fiction and Forecast.* London: University of London.

Gould, S. J. (1991). *Bully for Brontosaurus: Reflections in Natural History.* London: Hutchinson Radius.

Grice, P. (1957). Meaning. *The Philosophical Review, 66*, 377–388.

Grice, P. (1968). Utterer's meaning, sentence meaning and word meaning. *Foundations of Language, 4*, 225–242.

Grice, P. (1974). Method in philosophical psychology (from the banal to the bizarre). *Proceedings and Addresses of the American Philosophical Association, 48*, 23–53.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax & Semantics 3: Speech Acts.* (pp. 41–58).

Grice, P. (1981). Presupposition and conversational implicature. In P. Cole (Ed.), *Radical Pragmatics.* (pp. 183–198). New York: Academic Press.

Grice, P. (1989a). Logic and conversation. In *Studies in the Way of Words.* (pp. 22–40). Cambridge, Mass: Harvard University Press. (Originally published in P. Cole and J. Morgan, (Ed.s), 'Syntax and Semantics'. Academic Press, New York, 1975.)

Grice, P. (1989b). Retrospective Epilogue. In *Studies in the Way of Words.* (pp. 339–385). Cambridge, Mass: Harvard University Press.

Grice, P. (1989c). *Studies in the Way of Words.* Cambridge, Mass: Harvard University Press.

Grice, P. (2001). *Aspects of Reason.* Oxford: Clarendon Press.

Griggs, R. A. & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*(3), 407–420.

Griggs, R. A. & Cox, J. R. (1993). Permission schemas and the selection task. *The Quarterly Journal of Experimental Psychology: A. Human Experimental Psychology, 46*(4), 637–651.

Hacking, I. (1979). What is logic? *The Journal of Philosophy, 76*(6), 285–319.

Hare, R. M. (1952). *The Language of Morals.* Oxford: Clarendon Press.

Harman, G. (1984). Logic and reasoning. *Synthèse, 60*(1), 107–127.

Harman, G. (1986). *Change in View: Principles of Reasoning.* Cambridge, Mass: MIT Press.

Harman, G. (1999). *Reasoning, Meaning and Mind.* Oxford: Clarendon Press.

Harman, G. (2003). Aspects of reason II (Review of 'Aspects of Reason', Paul Grice, 2001). *The Philosophical Quarterly, 53*(211), 280–284.

Harman, G. (2004). Practical aspects of theoretical reasoning. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 45–56). New York: Oxford University Press.

Hauser, M. D. (2003). Knowing about knowing: Dissociations between perception and action systems over evolution and during development. *Annals of the New York Academy of Sciences, 1001*(1), 79–103.

Hausman, D. M. (2006). Philosophy of economics. (Article in The Stanford Encyclopedia of Philosophy, an online publication). Summer 2006 edition. Retrieved 2007/07/25, http://plato.stanford.edu/archives/sum2006/entries/economics/

Hayes, P. J. (1987). What the frame problem is and isn't. In Z. W. Pylyshyn (Ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence.* (pp. 123–138). Norwood, N.J: Ablex.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366–378.

Hertwig, R. & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275–305.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118*(2), 248–271.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications.* (pp. 11–42). Washington, DC: Georgetown University Press.

Horn, L. R. (2006). The border wars: A neo-Gricean perspective. In K. Heusinger & K. Turner (Eds.), *Where Semantics Meets Pragmatics.* (pp. 21–48). Oxford: Elsevier.

Horsey, R. S. (2006). *The Content and Acquisition of Lexical Concepts.* PhD thesis, University College London.

Hume, D. (2003). *A Treatise of Human Nature.* Mineola, N.Y: Dover Publications. (Originally published 1739.)

Jary, M. (2005). *Assertion and Mood: A Cognitive Account.* PhD thesis, University College London.

Johnson-Laird, P. N. (1975). Models of deduction. In R. J. Falmagne (Ed.), *Reasoning: Representation and Process in Children and Adults.* (pp. 7–54). Hillsdale, NJ: Lawrence Erlbaum.

Johnson-Laird, P. N. (1983). *Mental Models.* New York: Cambridge University Press.

Johnson-Laird, P. N. (1997a). An end to the controversy? A reply to Rips. *Minds and Machines, 7*(3), 425–432.

Johnson-Laird, P. N. (1997b). Rules and illusions: A critical study of Rips's 'The Psychology of Proof'. *Minds and Machines, 7*(3), 387–407.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology,* 109–110.

Johnson-Laird, P. N. (2004). Mental models and reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 169–204). Cambridge: Cambridge University Press.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction.* Hove: Erlbaum.

Johnson-Laird, P. N., Girotto, V. & Legrenzi, P. (2003). Mental models: a gentle guide for outsiders. *Web document.* Retrieved 2007-01-12, http:/

/www.si.umich.edu/ICOS/gentleintro.html

Johnson-Laird, P. N., Legrenzi, P., Girotto, V. & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science, 288*(5465), 531–532.

Johnson-Laird, P. N. & Savary, F. (1996). Illusory inferences about probabilities. *Acta Psychologica, 93*, 69–90.

Johnson-Laird, P. N. & Wason, P. C. (Eds.). (1977). *Thinking: Readings in Cognitive Science.* Cambridge: Cambridge University Press.

Jones, M. A. (2002). *Textbook on Torts* (8th ed.). Oxford: Oxford University Press.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. W. Griffin & D. Kahneman (Eds.), *Heuristics and biases: the psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*(2), 123–141.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–91; discusion 592–6.

Kamp, H. (2002). A theory of truth and semantic representation. In P. Portner & B. H. Partee (Eds.), *Formal Semantics: The Essential Readings.* (pp. 189–222). Oxford: Blackwell. (Originally published in Groenendijk, J. Janssen, T. & Stokhof, M. (Ed.s) 1981, 'Truth, Interpretation and Information', pp 1-41. Dordrecht: Foris.)

Kasher, A. (1976). Conversational maxims and rationality. In A. Kasher (Ed.), *Language in Focus: Foundations, Methods and Systems.* (pp. 197–216). Dordrecht, Holland: Reidel Publishing Company.

Kasher, A. (1982). Gricean inference revisited. *Philosophica, 29*, 25–44.

Kasher, A. (1991). Pragmatics and Chomsky's research program. In A. Kasher (Ed.), *The Chomskyan Turn.* (pp. 122–149). Oxford: Basil Blackwell.

Kasher, A. (no date). Rationality and pragmatics. Retrieved 2006-09-07, http:/ /www.tau.ac.il/~kasher/prprag.htm

Kenny, A. J. P. (1963). *Action, Emotion and Will.* London: Routledge & K. Paul.

Keynes, J. M. (1921). *A Treatise on Probability.* London: Macmillan.

Kroger, J. K., Cheng, P. W. & Holyoak, K. J. (1993). Evoking the permission schema: The impact of explicit negation and a violation-checking context. *The Quarterly Journal Of Experimental Psychology A: Human Experimental Psychology, 46*(4), 615–635.

Kubovy, M. & Epstein, W. (2002). Internalization: A metaphor we can live without. *Behavioral and Brain Sciences, 24*(4), 618–625.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*. (pp. 91–195). Cambridge: Cambridge University Press.

Land, S. K. (1974). *From Signs to Propositions: The Concept of Form in Eighteenth-Century Semantic Theory.* London: Longman.

Lavie, N. (2000). Selective attention and cognitive control: Dissociating attentional functions through different types of load. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII.* (pp. 175–194). Cambridge, Mass: MIT Press.

Lavie, N. (2006). The role of perceptual load in visual awareness. *Brain Research, 1080*(1), 91–100.

Lavie, N. & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perceptual Psychophysics, 56*(2), 183–197.

Lemmon, E. J. (1978). *Beginning Logic.* Indianapolis: Hackett.

Levinson, S. C. (2006). On the human 'interactional engine'. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction.* (pp. 39–69). Oxford: Berg.

Liszkowski, U. (2005). Human twelve-month-olds point cooperatively to share interest with and helpfully provide information for a communicative partner. *Gesture, 5*(1-2), 135–154.

Liszkowski, U. (2006). Infant pointing at twelve months: Communicative goals, motives, and social-cognitive abilities. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction.* (pp. 153–178). Oxford: Berg.

Liszkowski, U., Carpenter, M., Henning, A., Striano, T. & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science, 7*(3), 297–307.

Liszkowski, U., Carpenter, M., Striano, T. & Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. *Journal of Cognition and Development, 7*(2), 173–187.

Liszkowski, U., Carpenter, M. & Tomasello, M. (2007). Reference and attitude in infant pointing. *Journal of Child Language, 34*(1), 1–20.

López, F. J., Cobos, P. L., Caño, A. & Shanks, D. R. (undated). An associationist view of biases in causal and probabilistic judgment. *ELSE working papers, 003.*

Macbeth, D. (2005). *Frege's Logic.* Cambridge, MA: Harvard University Press.

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *The Quarterly Journal Of Experimental Psychology A. Human Experimental Psychology, 48*(1), 188–207.

Makinson, D. C. (1965). The paradox of the preface. *Analysis, 25*(6), 205–207.

Malle, B. F., Moses, L. J. & Baldwin, D. A. (2001). Introduction: The significance of intentionality. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition.* (pp. 1–26). Cambridge, Massachusetts: Bradford Books, MIT Press.

Manktelow, K. I. (1999). *Reasoning and Thinking.* Hove: Psychology Press.

Manktelow, K. I. & Over, D. E. (1993). *Rationality: Psychological and Philosophical Perspectives.* London: Routledge.

March, J. G. & Simon, H. A. (1958). *Organizations.* New York: Wiley.

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information.* San Francisco: Freeman.

Marsh, B., Todd, P. M. & Gigerenzer, G. (2004). Cognitive heuristics: Reasoning the fast and frugal way. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 273–287). Cambridge: Cambridge University Press.

Martin, E. A. (2002). *A Dictionary of Law* (5th ed.). Oxford: Oxford University Press.

Matsui, T. (2000). *Bridging and Relevance.* Amsterdam: J. Benjamins.

McClennen, E. F. (1990). *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge: Cambridge University Press.

McGilvray, J. (2005). Meaning and creativity. In J. McGilvray (Ed.), *The Cambridge Companion to Chomsky.* (pp. 204–222). Cambridge: Cambridge University Press.

McLeod, P. & Dienes, Z. (1996). Do fielders know where to go to catch the ball or only how to get there? *Journal of Experimental Psychology: Human Perception and Performance, 22*(3), 531–543.

McLeod, P., Reed, N. & Dienes, Z. (2001). Toward a unified fielder theory: What we do not yet know about how people run to catch a ball. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 1347–1355.

McLeod, P., Reed, N. & Dienes, Z. (2003). How fielders arrive in time to catch the ball. *Nature, 426*, 244–245.

McLeod, P., Reed, N. & Dienes, Z. (2006). The generalized optic acceleration cancellation theory of catching. *Journal of Experimental Psychology: Human Perception & Performance, 32*(1), 139–148.

Mele, A. R. (1997a). Introduction. In A. R. Mele (Ed.), *The Philosophy of Action.* (pp. 1–26). Oxford: Oxford University Press.

Mele, A. R. (Ed.). (1997b). *The Philosophy of Action.* Oxford: Oxford University Press.

Mele, A. R. (2001). Acting intentionally: Probing folk notions. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition.* (pp. 27–44). Cambridge, Massachusetts: Bradford Books, MIT Press.

Mele, A. R. & Rawling, P. (2004a). Introduction: Aspects of rationality. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 3–16). New York: Oxford University Press.

Mele, A. R. & Rawling, P. (Eds.). (2004b). *The Oxford Handbook of Rationality.* New York: Oxford University Press.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology, 24*(4), 470–476.

Meltzoff, A. N. & Brooks, R. (2001). "Like me" as a building block for understanding other minds: Bodily acts, attention and intention. In B. F.

Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition.* (pp. 171–192). Cambridge, Mass: Bradford Books, MIT Press.

Mill, J. S. (1856). *System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation.* London: John W. Parker.

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories.* Cambridge, Mass: MIT Press.

Millikan, R. G. (1987). What Peter thinks when he hears Mary speak (Reply to Sperber and Wilson, Précis of 'Relevance: Communication and Cognition'). *Behavioral and Brain Sciences, 10*, 725–726.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice.* Cambridge, Mass: MIT Press.

Millikan, R. G. (2003). In defense of public language. In L. M. Antony & N. Hornstein (Eds.), *Chomsky and His Critics.* (pp. 215–237). Oxford: Blackwell.

Millikan, R. G. (2005). Semantics/pragmatics: Purposes and cross-purposes. In *Language: A Biological Model.* (pp. 187–220). Oxford: Oxford University Press.

Mitchell, P., Robinson, E. J. & Thompson, D. E. (1999). Children's understanding that utterances emanate from minds: Using speaker belief to aid interpretation. *Cognition, 72*(1), 45–66.

Morton, A. (2006). Review of 'Understanding People: Normativity and Rationalizing Explanation', Alan Millar, 2004. *Mind, 115*(459), 777–780.

Murphy, D. (2006). On Fodor's analogy: Why psychology is like philosophy of science after all. *Mind & Language, 21*(5), 553–564.

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind & Language, 21*(5), 565–586.

Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy, 15*(5), 509–559.

Neeleman, A. & van de Koot, H. (2004). The Grammatical Code. *Ms. UCL.*

Nehemas, A. (1994). Meno's Paradox and Socrates as a teacher. In J. M. Day (Ed.), *Plato's Meno In Focus.* (pp. 221–248). London: Routledge.

Newell, A. & Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery, 19*(3), 113–126.

Newstead, S. E., Pollard, P., Evans, J. S. B. T. & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45*(3), 257–284.

Nickerson, R. S. (2004). Teaching reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 410–442). Cambridge: Cambridge University Press.

Nicolle, S. & Clark, B. (1999). Experimental pragmatics and what is said: A response to Gibbs and Moise. *Cognition, 69*(3), 337–354.

Nietzsche, F. W. (1968). *Basic Writings of Nietzsche.* New York: Modern Library.

Nisbett, R. E. & Borgida, E. (1975). Attribution and the psychology of prediction.

Journal of Personality and Social Psychology, 32(5), 932–943.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*(2), 165–188.

Noveck, I. A. & O'Brien, D. P. (1996). To what extent do pragmatic reasoning schemas affect performance on Wason's selection task? *The Quarterly Journal of Experimental Psychology: Section A, 49*(2), 463–489.

Noveck, I. A. & Prado, J. (2007). Intelligence and reasoning are not one and the same. Commentary on Jung, R. E. & Haier, R. J. 'The Parieto-Frontal Integration Theory (P-FIT) of intelligence: Converging neuroimaging evidence'. *Behavioral and Brain Sciences, 30*(2), 163–164.

Noveck, I. A. & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (Ed.), *Pragmatics.* (pp. 184–212). Basingstoke: Palgrave Macmillan.

Nozick, R. (1973). Distributive justice. *Philosophy and Public Affairs, 3*(1), 45–126.

Nozick, R. (1974). *Anarchy, State, and Utopia.* New York: Basic Books.

Nuti, M. (2003). *Ethnoscience: Examining Common Sense.* PhD thesis, University College London.

O'Brien, D. P. (2004). Mental-logic theory: What it proposes, and reasons to take this proposal seriously. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 205–233). Cambridge: Cambridge University Press.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659–677.

Oaksford, M. & Chater, N. (1993). Reasoning theories and bounded rationality. In K. I. Manktelow & D. Over (Eds.), *Rationality: Psychological and Philosophical Perspectives.* (pp. 31–60). London: Routledge.

Onishi, K. H. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Origgi, G. & Sperber, D. (2000). Evolution, communication, and the proper function of language. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the Human Mind: Modularity, Language and Meta-Cognition.* (pp. 140–169). Cambridge: Cambridge University Press.

Osherson, D. N. (1975a). Logic and models of logical thinking. In R. Falmagne (Ed.), *Reasoning: Representation and process in children and adults.* (pp. 81–91). Hillsdale, N.J.: Erlbaum.

Osherson, D. N. (1975b). *Reasoning in Adolescence: Deductive Inference.* Hillsdale, N.J.: L. Erlbaum Associates.

Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy, 14*, 473–513.

Parikh, P. (2001). *The Use of Language.* Stanford, California: CSLI Publications.

Pashler, H. E. (1998). *The Psychology of Attention.* Cambridge, Mass: MIT Press.

Perner, J. & Lopez, A. (1997). Children's understanding of belief and disconfirming visual evidence. *Cognitive Development, 12*, 463–475.

Piattelli-Palmarini, M. (1994). *Inevitable Illusions: How Mistakes of Reason Rule*

Our Minds. Chichester: Wiley.

Pinker, S. (2005). So how does the mind work? *Mind & Language, 20*(1), 1–24.

Plato. (1991). *The Republic: The Complete and Unabridged Jowett Translation.* New York: Vintage Books.

Poe, E. A. (1841). The Murders in the Rue Morgue. Retrieved 15-01-2007, http://en.wikisource.org/wiki/The_Murders_in_the_Rue_Morgue

Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*(6035), 314–319.

Poletiek, F. (2001). *Hypothesis Testing Behavior.* Philadelphia: Psychology Press.

Politzer, G. (1990). Immediate deduction between quantified sentences. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie & G. Erdos (Eds.), *Lines of Thinking: Reflections on the Psychology of Thought.* (pp. 85–97). London: John Wiley.

Politzer, G. (2004). Reasoning, judgement and pragmatics. In I. A. Noveck & D. Sperber (Eds.), *Experimental Pragmatics.* (pp. 94–115). London: Palgrave Macmillan.

Politzer, G. (2005). Uncertainty and the suppression of inferences. *Thinking & Reasoning, 11*(1), 5–33.

Politzer, G. & Bourmaud, G. (2002). Deductive reasoning from uncertain premises. *British Journal of Psychology, 93*(3), 345–381.

Politzer, G. & Macchi, L. (2005). The representation of the task: The case of the Lawyer-Engineer problem. In *The Shape of Reason. Essays in Honour of P. Legrenzi.* (pp. 119–135). Hove: Psychology Press.

Politzer, G. & Nguyen-Xuan, A. (1992). Reasoning about conditional promises and warnings: Darwinian algorithms, mental models, relevance judgements or pragmatic schemas? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 44*(3), 401–421.

Politzer, G. & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research, 20*(2), 83–103.

Popper, K. R. (1959). *The Logic of Scientific Discovery.* London: Hutchinson.

Pouscoulous, N. & Noveck, I. (2004). Investigating scalar implicature, abstract for workshop on implicature and conversational meaning, 16-20 August, Nancy. 2005-08-04 edition. Retrieved 2005-08-04, http://www.ru.nl/filosofie/implicatures/abstracts/pouscoulous.pdf

Progovac, L., Paesani, K., Casielles, E., & Barton, E. (Eds.). (2006). *The Syntax of Nonsententials: Multidisciplinary Perspectives.* Amsterdam: John Benjamins.

Rawls, J. (1993). *Political Liberalism.* New York: Columbia University Press.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious.* Oxford: Clarendon Press.

Recanati, F. (1989). The pragmatics of what is said. *Mind & Language, 4,* 295–329.

Recanati, F. (2002a). Unarticulated constituents. *Linguistics and Philosophy, 25*(3), 299–345.

Recanati, F. (2002b). Does linguistic communication rest on inference? *Mind & Language, 17,* 105–126.

Recanati, F. (2004). *Literal Meaning.* Cambridge: Cambridge University Press.

Reid, T. (1855). *Essays on the Intellectual Powers of Man* (6th ed.). Boston, Mass: Phillips, Sampson and Company. (Originally published 1785.)

Reid, T. (2004). Essays on the Active Powers of the Human Mind, edited by Jonathan F. Bennett. Retrieved 2-09-2007, http:/ /www.earlymoderntexts.com/pdfbig/reidabig.pdf (Originally published 1788.)

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*(1), 38–71.

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking.* Cambridge, Mass: MIT Press.

Rips, L. J. (1997). Goals for a theory of deduction: Reply to Johnson-Laird. *Minds and Machines, 7*(3), 409–424.

Roberts, M. J. (2004). Heuristics and reasoning I: Making deduction simple. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning.* (pp. 234–272). Cambridge: Cambridge University Press.

Roberts, M. J. & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *The Quarterly Journal of Experimental Psychology A, 54,* 1031–1048.

Roberts, M. J. & Newton, E. J. (2003). Individual differences in the development of reasoning strategies. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological Perspectives On Reasoning, Judgment, and Decision Making.* (pp. 23–44). Hoboken, NJ: Wiley.

Rockwell, T. (2005). Attractor spaces as modules: A semi-eliminative reduction of symbolic AI to dynamic systems theory. *Minds and Machines, 15*(1), 23–55.

Ruffman, T. & Perner, J. (2005). Do infants really understand false belief? Response to Leslie. *Trends in Cognitive Science, 9*(10), 462–463.

Russell, B. (1983). Am I an atheist or an agnostic? In K. Blackwell (Ed.), *The collected papers of Bertrand Russell.* (McMaster University ed., pp. 89–92). London: G. Allen & Unwin.

Ryan, S. (1991). The preface paradox. *Philosophical Studies, 64*(3), 293–307.

Ryle, G. (1949). *The Concept of Mind.* London: Hutchinson's University Library.

Samuels, R. & Stich, S. (2004). Rationality and psychology. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 279–300). Oxford: Oxford University Press.

Samuels, R., Stich, S. & Bishop, M. (2002). Ending the rationality wars: How to make disputes about human rationality disappear. In R. Elio (Ed.), *Common Sense, Reasoning, and Rationality.* (pp. 236–268). New York: Oxford University Press.

Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics: The Arne Ryde Memorial Lectures.* Oxford: Clarendon Press.

Saul, J. M. (2002a). Speaker meaning, what is said, and what is implicated. *Noûs, 36*(2), 228–248.

Saul, J. M. (2002b). What is said and psychological reality; Grice's project and relevance theorists' criticisms. *Linguistics and Philosophy, 25*(3), 347–372.

Schaeken, W., De Vooght, G., Vandierendonck, A., & D'Ydewalle, G. (Eds.). (2000). *Deductive Reasoning and Strategies.* Mahwah, N.J: L. Erlbaum Associates.

Schiffer, S. R. (1972). *Meaning.* Oxford: Oxford University Press.

Schlosser, M., E. (2007). Basic deviance reconsidered. *Analysis, 67*(295), 186–194.

Schueler, G. F. (2001). Action explanations: Causes and purposes. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition.* (pp. 251–264). Cambridge, Massachusetts: Bradford Books, MIT Press.

Schwarz, N., Strack, F., Hilton, D. & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of irrelevant information. *Social Cognition, 9*(1), 67–84.

Scott, J. (2000). Rational choice theory. In G. K. Browning, A. Halcli & F. Webster (Eds.), *Understanding Contemporary Society: Theories of the Present.* (pp. 126–138). London: Routledge and K. Paul.

Segal, G. (1996). The modularity of theory of mind. In P. Carruthers & P. Smith (Eds.), *Theories of Theories of Mind.* (pp. 141–157). Cambridge: Cambridge University Press.

Sen, A. K. (1976). Liberty, unanimity and rights. *Economica, 43*(171), 217–245.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs, 6*(4), 317–344.

Sen, A. K. (1993). Internal consistency of choice. *Econometrica, 61*(3), 495–521.

Shafir, E. & Leboeuf, R. A. (2002). Rationality. *Annual Review of Psychology,* 491–518.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Sibley, W. M. (1953). The rational versus the reasonable. *The Philosophical Review, 62*(4), 554–560.

Sides, A., Osherson, D., Bonini, N. & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition, 30*(2), 191–198.

Simon, H. A. (1947). *Administrative Behavior.* New York: Macmillan.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.

Simon, H. A. (1957a). *Administrative Behavior; a Study of Decision-Making Processes in Administrative Organization* (2nd ed.). New York: Macmillan.

Simon, H. A. (1957b). *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in a Social Setting.* New York: Wiley.

Simon, H. A. (1962). The architecture of complexity: Hierarchic systems. *Proceedings of the American Philosophical Society, 106,* 467–482.

Simon, H. A. (1969). *The Sciences of the Artificial* (Karl Taylor Compton lectures, 1968). Cambridge, Mass: MIT Press.

Simon, H. A. (1982). *Models of Bounded Rationality.* Cambridge, Mass: MIT

Press.

Simon, H. A. (1983). *Reason in Human Affairs* (Harry Camp lectures at Stanford University: 1982). Stanford, Calif: Stanford University Press.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*(1), 1–19.

Simon, H. A. (1997). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (4th ed.). New York: Free Press.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society, 1*(1), 25–39.

Simon, H. A. & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research, 6*(1), 1–10.

Simpson, J. A. & Weiner, E. S. C. (1991). *The Compact Oxford English Dictionary.* Oxford: Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Smith, M. (2004). Humean rationality. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 75–92). New York: Oxford University Press.

Smith, N. V. (1999). *Chomsky: Ideas and Ideals.* Cambridge: Cambridge University Press.

Smith, N. V. & Tsimpli, I.-M. (1995). *The Mind of a Savant: Language Learning and Modularity.* Oxford: Blackwell.

Sorensen, R. (2004). Paradoxes of rationality. In A. R. Mele & P. Rawling (Eds.), *The Oxford Handbook of Rationality.* (pp. 257–275). New York: Oxford University Press.

Southgate, V., Senju, A. & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.

Southgate, V., van Maanen, C. & Csibra, G. (2007). Infant pointing: communication to cooperate or communication to learn? *Child Development, 78*(3), 735–740.

Sperber, D. (1994). Understanding verbal understanding. In *What is Intelligence?* (pp. 179–198). Cambridge: Cambridge University Press.

Sperber, D. (1995). How do we communicate? In J. Brockman & K. Matson (Eds.), *How Things Are: A Science Toolkit to the Mind.* (pp. 191–199). New York: W. Morrow.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language, 12*(1), 67–83.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In *Metarepresentations: A Multidisciplinary Perspective.* (pp. 117-137). New York: Oxford University Press.

Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics, 29*, 401–403.

Sperber, D., Cara, F. & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*(1), 31–95.

Sperber, D. & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. *Cognition, 85*(3), 277–290.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and Cognition* (2nd ed.). Oxford: Blackwell. (Originally published 1986: page references are to 2nd ed., 1995.)

Sperber, D. & Wilson, D. (1987a). Authors' response: Presumptions of relevance (Response to various authors' comments on Précis of 'Relevance: Communication and Cognition'). *Behavioral and Brain Sciences, 10*, 736–754.

Sperber, D. & Wilson, D. (1987b). Précis of 'Relevance: Communication and Cognition'. *Behavioral and Brain Sciences, 10*, 697–754.

Sperber, D. & Wilson, D. (1995). Postface. In *Relevance: Communication and Cognition*. (2nd ed., pp. 255–279). Oxford: Blackwell.

Sperber, D. & Wilson, D. (1996). Fodor's frame problem and relevance theory (reply to Chiappe & Kukla). *Behavioral and Brain Sciences, 19*(3), 530–532.

Sperber, D. & Wilson, D. (1998). The mapping between the mental and the public lexicon. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary Themes*. (pp. 184–200). Cambridge University Press.

Sperber, D. & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language, 17*(1–2), 3–23.

Sperber, D. & Wilson, D. (2004). Relevance theory. In L. R. Horn & G. L. Ward (Eds.), *The handbook of pragmatics*. (pp. 607–632). Malden, MA: Blackwell.

Sperber, D. & Wilson, D. (2005). Pragmatics. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy*. (pp. 468–501). Oxford: Oxford University Press.

Stanovich, K. E. (1999). *Who is Rational? : Studies of Individual Differences in Reasoning*. Mahwah, N.J: Lawrence Erlbaum Associates.

Stanovich, K. E. & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking & Reasoning, 4*(4), 289–317.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665.

Stanovich, K. E. & West, R. F. (2004). The rationality debate as a progressive research program. *Behavioral and Brain Sciences, 26*(04), 531–533.

Stenning, K. & Monaghan, P. (2004). Strategies and knowledge representation. In J. P. Leighton & R. J. Sternberg (Eds.), *The Nature of Reasoning*. (pp. 129–168). Cambridge: Cambridge University Press.

Stigler, G. J. (1961). The economics of information. *The Journal of Political Economy, 69*(3), 213–225.

Styles, E. A. (1997). *The Psychology of Attention*. Hove, East Sussex, UK: Psychology Press.

Surian, L., Caldi, S. & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.

Teigen, K. H. (1974). Overestimation of subjective probabilities. *Scandinavian Journal of Psychology, 15*, 56–62.

Templeton, L. M. & Wilcox, S. (2000). A tale of two representations: The misinformation effect and children's developing theory of mind. *Child Development, 71*, 402–416.

Tentori, K., Bonini, N. & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science, 28*(3), 467–477.

Todd, P. M. & Gigerenzer, G. (2000). Précis of 'Simple heuristics that make us smart'. *Behavioral & Brain Sciences, 23*(5), 727–41; discussion 742–80.

Todd, P. M. & Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of Economic Psychology, 24(2)*(2), 143–165.

Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–91; discussion 691–735.

Tomasello, M., Carpenter, M. & Liszkowski, U. (2007). A new look at infant pointing. *Child Development, 78*(3), 705–722.

Tsimpli, I. M. & Smith, N. (1998). Modules and quasi-modules: Language and theory of mind in a polyglot savant. *Learning and Individual Differences, 10*(3), 193–215.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Uchida, H. (2007). Logic in pragmatics. *UCL Working Papers in Linguistics, 19*, 285–318.

van Dalen, D. (2004). *Logic and Structure* (4th ed.). Berlin: Springer-Verlag.

van der Henst, J. B. (2006). "Symposium on Cognition and Rationality: Part I" Relevance effects in reasoning. *Mind & Society, 5*(2), 229–245.

van der Henst, J. B., Politzer, G. & Sperber, D. (2002). When is a conclusion worth deriving? A relevance-based analysis of indeterminate relational problems. *Thinking & Reasoning, 8*(1), 1–20.

van Oostendorp, H. & De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta psychologica, 74*(1), 35-46.

van Oostendorp, H. & Kok, I. (1990). Failing to notice errors in sentences. *Language and Cognitive Processes, 5*(2), 105-113.

von Helmholtz, H. (1962). *Treatise on Physiological Optics. [Handbuch Der Physiologischen Optik.]* (J. P. C. Southall, Trans.). Dover. (Originally published 1925.)

von Wright, G. H. (1971). *Explanation and Understanding.* London: Routledge and K. Paul.

Vriend, N. J. (1996). Rational behavior and economic theory. *Journal of Economic Behavior and Organization, 29*(2), 263–285.

Walker, D. P. (1972). Leibniz and language. *Journal of the Warburg and Courtauld Institutes, 35*, 294–307.

Wallace, C. M. (1980). Coleridge's theory of language. *Philological Quarterly, 59*, 338–352.

Warner, R. (2001). Introduction: Grice on reasons and rationality. In R. Warner (Ed.), *Aspects of Reason.* (pp. vii–xxxviii). Oxford: Clarendon Press.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology.* (pp. 135–151). Harmondsworth, Middlesex: Penguin.

Wason, P. C. (1968a). On the failure to eliminate hypotheses: A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking And Reasoning: Selected Readings.* Baltimore: Penguin.

Wason, P. C. (1968b). Reasoning about a rule. *The Quarterly Journal Of Experimental Psychology, 20*(3), 273–281.

Wason, P. C. & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23,* 63–71.

Wedgwood, R. (2006). The normative force of reasoning. *Noûs, 40*(4), 660–686.

Wellman, H. M. (1990). *The Child's Theory of Mind.* Cambridge, Mass: MIT Press.

Wellman, H. M., Cross, D. & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14,* 246–249.

Wharton, T. (2002). Paul Grice, saying and meaning. *UCL Working Papers in Linguistics, 14,* 207–248.

Wharton, T. (2003). *Pragmatics and the 'Showing/Saying' Continuum.* PhD thesis, University College London.

Wilson, D. (2000). Metarepresentations in linguistic communication. In D. Sperber (Ed.), *Metarepresentations: a multidisciplinary perspective.* (pp. 411–448). Oxford ; New York: Oxford University Press.

Wilson, D. (2005). New directions for research on pragmatics and modularity. *Lingua, 115*(8), 1129–1146.

Wilson, D. & Matsui, T. (1998). Recent approaches to bridging: Truth, coherence, relevance. *UCL Working Papers in Linguistics, 10,* 173–200.

Wilson, D. & Sperber, D. (1981). On Grice's theory of conversation. In P. Werth (Ed.), *Conversation and Discourse.* (pp. 155–178). London: Croom Helm.

Wilson, D. & Sperber, D. (2002). Truthfulness and relevance. *Mind, 111*(443), 583–632.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.

Winch, P. (1958). *The Idea of a Social Science and Its Relation to Philosophy.* London: Routledge & Paul Humanities Press.

Winner, E. (1988). *The Point of Words: Children's Understanding of Metaphor and Irony.* Cambridge, Mass: Harvard University Press.

Winner, E., Engel, M. & Gardner, H. (1980). Misunderstanding metaphor:

What's the problem? *Journal of Experimental Child Psychology, 30*(1), 22–32.

Winner, E., Rosentiel, A. & Gardner, H. (1976). The development of metaphoric understanding. *Developmental Psychology,* 289–297.

Winter, S. G. (1975). Optimization and evolution in the theory of the firm. In R. H. Day & T. Groves (Eds.), *Adaptive Economic Models.* (pp. 73–118). New York: Academic Press.

Woodward, A. A., Sommerville, J. A. & Guajardo, J. J. (2001). How infants make sense of intentional action. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition.* (pp. 149–170). Cambridge, Mass: Bradford Books, MIT Press.

This thesis is set in 11 point Warnock Pro.

The writing and setting were both done in *Mellel* on an Apple Mac. Chapters 4 and 5 and part of chapter 1 were drafted in *Scrivener*. The bibliography was generated using *Bookends*.