PHILOSOPHY OF MIND

Seeing Through Self-Deception
By ANNETTE BARNES
Cambridge University Press, 1998. x + 182 pp. £32.50

Annette Barnes has provided us with a sophisticated new treatment of self-deception. However, the book is not an easy read. It requires the reader to remain alert to the nuances in the line of argument. It also could have done with a little rewriting in places. There are a number of long footnotes that involve commentary on other peoples' work or related matters which should have been integrated into the main text or abandoned altogether. These facts make the book a not-altogether-ideal presentation of her views. This is a shame since it repays study.

Barnes begins by considering the connection between self-deception and other-deception. Some hold that self-deception should not be modelled on other-deception because this gives rise to paradoxes. First, there is the *doxastic paradox*. Self-deception involves a subject in believing something he or she knows or truly believes is false. Second there is the *strategic paradox*. Self-deception involves a subject being deceived into believing that p as a result of his or her own duplicitous intention. Barnes claims that modelling self-deception on other-deception only involves a version of the second of these paradoxes. That is because

> It need not be the case, when A intentionally deceives B into believing that p, that (a) A knows or truly believes that something is false, and (b) A intentionally gets B to believe that is true (p. 5).

She invites us to consider the following case to illustrate the point. A knows that there is a "shy, unpredictable" rabbit in B's garden—B being a five year old child. She wants B to see a rabbit so she places a model rabbit in the garden. Later on, B tells A that she (B) has seen a rabbit. A is unsure whether B saw the model rabbit or the real rabbit. If B saw the model rabbit, she was deceived by A. However, Barnes claims, there is no proposition that A knows or truly believes is false which B believes to be true (pp. 11–12). Hence a necessary condition for the first paradox isn't met.

It is not clear that this case illustrates what Barnes wants. There does seem to be something which A knows or truly believes is false that B could reasonably be said to believe, namely that everything which looks like a rabbit in the garden is a rabbit (cf. pp. 12–13). Arguably, B wouldn't have concluded that there is a rabbit in the garden if she had believed that it is false that everything which looks like a rabbit in the garden is a rabbit. This might be enough to attribute to B a belief that everything which looks like a rabbit in the garden is a rabbit. In which case, Barnes's example collapses. Barnes may argue that it is not appropriate to attribute this belief to a 5 year old. But this would show an error in Barnes's preliminary characterisation of the standard picture of other-deception. Other-deception can occur if the deceiver knows (or believes truly) that something is false and intentionally gets the deceived *not to have the belief* that the thing in question is false (which is not necessarily to believe that it is true). Modelling self-deception on this kind of other-deception involves an even more serious though related paradox to the doxastic paradox: the contradictory ascription of belief. Either way, Barnes doesn't seem to get the relationship between self-deception and other-deception quite right.

Barnes's own theory is a development of Mark Johnston's. She joins him in rejecting the idea that self-deception is intentional. Her grounds for this did not seem conclusive. She argues that intentions must be non-inferentially recognisable (p. 89). Initially, it might look as if this makes the strategic paradox insurmountable. On reflection, it is not so obvious. Of course, while a subject is self-deceived, he or she will not have non-inferentially *accessed* his or her intention to deceive him or herself. This doesn't mean that, if he or she is prompted to think about the issue, he or she will fail to non-inferentially recognise that he or she has been intending to bias themselves in such and such a way. If this is right, then the self-deceptive intention may be non-inferentially *accessible* to the self-deceived.

Barnes takes the ascription of an intention to a subject to imply his or her susceptibility to entertain and answer a certain kind of why question: what was your reason? What was your intention? (pp. 92–93). However, she does not explain why one should require this of every single intention ascribed to an individual. Intentions have a causal-rational role in the explanation of agents' behaviour. Suppose there is a state of an agent that plays the same

causal-rational role except that the agent is unable to respond to the appropriate why question. I think it is reasonable to classify this state as an intention even if it is not accessible in the way indicated.

Barnes also agrees with Johnston that self-deception is purposive. It involves a mental mechanism whose function is to reduce anxiety (pp. 31–32). Her full analysis of self-deception (or what she calls self-deceiving oneself) runs as follows

One self-deceives oneself into believing that p if and only if

(1) One has an anxious desire that q which causes one to be biased in favour of beliefs that reduce one's anxiety that not-q. This bias or partiality operating in one's acting or thinking or judging or perceiving etc. causes in the right way one to believe that p.

(2) The purpose of one's believing that p is to reduce one's anxiety that not-q.

(3) One is not intentionally biased or partial.

(4) One fails to make a high enough estimate of the causal role that one's anxious desire that q plays in one's acquiring the belief that p. One believes (wrongly, when condition 1 is met) that one's belief that p is justified (p. 117).

One is self-deceived in believing that p if p is (in addition) false (p. 118). Barnes claims that, if the purpose of a belief that p is to reduce anxiety, it does so *in itself* (p. 117, fn. 9). So all self-deception involves anxiety reduction. This does not seem to be right. Something may have as its purpose anxiety reduction and yet fail or only achieve this by producing something else which reduces anxiety. If I am right, then the account does not seem to provide a sufficient condition for self-deception. Take Davidson's famous case of the man who deliberately writes down the date of a meeting in his diary wrongly because he is anxious over meeting an enemy there. He knows that by the time of the meeting (some months away) he will forget what he has done and so miss the meeting. Barnes claims that this is not a case of self-deception because, since the belief itself fails to reduce anxiety, the purpose of the belief is not to reduce his anxiety (p. 113). This seems wrong. Purposes can fail to be carried out. Barnes may try to deal with this case by holding that the belief is brought about in the wrong way. So the case fails clause (1). However, this may rule out some cases which count as self-deception. It depends how the 'right way' is characterised.

Barnes's theory differs from Johnston's in the characterisation of how the content of the self-deceptive belief relates to the agent's motivation (see clause (2)). Johnston claims that self-deceptive beliefs are wishful beliefs, they are beliefs subjects desire to have as a result of which anxiety is reduced. However, this approach has difficulties with cases in which the self-deceived belief looks to be unwanted, for example, the jealous husband's belief that his wife is unfaithful (p. 35). Barnes suggests that the belief that p can be the result of a desire that q and serve to reduce an anxiety that not-q (where $q \neq p$). In such cases, the subject believes that if p then (probably) q. By believing that p the subject reduces his or her anxiety about some matter that the agent

believes to be related (p. 36). (It is assumed that q will be desired for its own sake. Otherwise the proposal gets a little more complicated (p. 39).)

Consider the case of the husband believing that his wife is unfaithful. The husband believes that his wife is unfaithful with his best friend because he wants to reduce his anxiety over the fact that he has strong feelings of jealousy which he views as blameworthy. He believes that if his wife were unfaithful with his best friend, then he would not be blameworthy for feeling jealous (pp. 43–44). Barnes denies that in such a case it is always right to attribute to the husband the desire that his wife is unfaithful (pp. 48–49). Rather the states which threaten to give rise to such a desire start "a complex causal non-intentional process" giving rise to the self-deceptive belief (p. 50). This seems right. However, I am not convinced that all cases of self-deception need to take even this slightly looser form. Some emotions seem just as purposeful as the anxiety-desire complex. A very jealous person takes the world to involve continual challenges to their loved one's virtue and so are ever on the look out. It does not take too much imagination to suppose that jealousy might have the purpose of increasing the likelihood of retaining one's mate. Another case is that of anger. Barnes notes that if Lucinda is angry with Ludwig for failing to invite her to a party, she may well judge Ludwig to have negative character traits on the basis of slight evidence or evidence to the contrary. She claims that this would not be self-deception because there is no belief produced with the function of reducing anxiety (p. 126). It is not clear to me that Barnes is right. I think we do say that someone who is very angry with someone else and accordingly forms a poor view of them is deceiving themselves. Perhaps that's because the purpose of anger is to enable us to assert ourselves in the face of harm done by others. A belief that the object of our anger has negative traits may well serve this purpose. A purposeful bringing about of a belief which itself has a certain function, when it is accompanied by an underestimation of the role of the emotion in the production of the belief, seems reasonably thought of as a case of self-deception whether it is rooted in anxiety, jealousy, anger or the like.

Barnes denies that self-deception involves the self-deceiver in believing that the total evidence favours the opposite belief to that produced by self-deception. Indeed, she claims that, if a subject becomes self-deceived, he or she conceives of the evidence as having changed in favour of the belief produced by self-deception (p. 147). All the self-deceived person need hold is that the belief produced by self-deception *may* be false (p. 146). A problem with this position is that one is not normally anxious that p unless one thinks that p is very likely to be true. I am not anxious that I might grow three heads. So the mere thought that the object of anxiety is possible cannot be enough to generate the elaborate strategies involved in self-deception. I don't doubt that one can think it is very likely that p without believing that the totality of evidence favours it. Nevertheless, believing that the totality of evidence favours p is one very obvious foundation for the anxiety that p. So I think that the belief about evidence is closer to the heart of some self-deception than Barnes allows. This point is quite compatible with allowing that self-deceivers are also adept at finding reasons for the belief that is

the product of self-deception (contrary to what Barnes seems to think, pp. 147–148).

[Typographical errors: p. 34, fn. 2: 'in' should read 'is'; p. 42, l. 5: 'George' should be 'John'; p. 42, l. 22: 'George' should be 'John'; p. 44, l. 29: the conditional should read 'if not-r then not-q'.]

THE UNIVERSITY OF NOTTINGHAM PAUL NOORDHOF

What Minds Can Do: Intentionality in a Non-Intentional World
By PIERRE JACOB
Cambridge University Press, 1997. xii + 300 pp. £40.00 cloth, £14.95 paper

In this book, Jacob attempts to defend intentional realism and physicalism despite the difficulties which this type of position faces. The two main difficulties are giving a naturalistic account of intentionality (which, for Jacob, amounts to showing how semantic properties of mental states supervene on some physical properties, such as the brain and the environment) and showing how propositional attitudes, in particular the semantic properties of propositional attitudes, can play a causal role in intentional behaviour. The structure of the book neatly mirrors this, the first half being a defence of information-based teleosemantics and the second half defending the thesis that broad content is not epiphenomenal.

Jacob presents in a detailed manner the basics of informational semantics, relying heavily on Dretske's account in *Knowledge and the Flow of Information*. He examines issues such as distinguishing normal or channel conditions from relevant information and the difference between conceptual or digital, propositional attitude content and non-conceptual or analogue, experiential content. He also makes some attempt to show how intentionality can have some bearing on consciousness; however, there is little new work here and his account relies both on the higher-order theory of consciousness and Evans's claim that to be conscious one must have concept-forming abilities.

Jacob then looks at several problems that face informational semantics. He outlines these and points the way towards solutions. His main preoccupation is rightly with the problem of ensuring that content is determinate when informational semantics threatens to render it indeterminate. He claims it can only be solved by turning to teleology. Rather than a benefit-based, non-informational account of teleology such as that forwarded by Millikan, Jacob offers an informationally-based, stimulus-based account. He argues quite convincingly that Millikan's account is not fully naturalistic, does not solve the indeterminacy problem and does not account for the special environments of some creatures. He then considers Fodor's objections to teleological solutions to the indeterminacy problem. While I was not quite convinced by Jacob's solution, the arguments and issues are clearly and fairly set out in this area. Indeed, it is a merit of the book in general that the author tries to provide a clear sense of the power and effectiveness of his own arguments both in relation to the problems themselves and in relation to the work of others.

The second half of the book begins with a discussion of the Computational Representational Theory of Mind. Jacob accepts the Language of Thought Hypothesis, as it accounts for the compositionality of semantic properties. He argues that it is not an empirical hypothesis but something conceptually necessary for thought. He also discusses the nature of psychological explanation, arguing against Davidson that non-strict laws can be causal laws.

Jacob discusses the view that meaning holism threatens the possibility of there being psychological laws. He argues persuasively that there is much wrong with the standard views in this area and that so long as one holds that psychological laws do not refer to the contents of psychological states but merely quantify over them, then one can vindicate psychology.

Jacob then examines how semantic properties or content could cause behaviour. The first problem is that of pre-emption. If semantic properties are higher-order properties of the brain and causation takes place on the physical level then there is a worry that mental processes look like mere pseudo-processes, not involved at the causal, physical level. Jacob points to a solution, based on Jackson and Pettit's work to the effect that a property can enter into a causal explanation without being directly causally efficacious. It can do so by featuring in a program explanation. Jacob, however, holds that this solution works only for narrow content—content that supervenes on the physical properties of the brain. This solution would therefore leave broad content epiphenomenal. Combined with his attack on theories of narrow content, the onus is thus on Jacob to show how it can be that broad content can cause intentional behaviour. The problem is that an externalist approach to content implies that semantic properties are non-local properties while the cause of an individual's behaviour is a local physical process. Jacob's solution rests on Dretske's componential view of behaviour, namely, that behaviour is a process whereby a propositional attitude causes some bodily movement. Behaviour is, therefore, not mere bodily movement and is not caused by propositional attitudes. Jacob argues that semantic properties can be structuring (as opposed to triggering) causes of the behavioural process. He concludes with an interesting discussion of the difference between the semantic properties of propositional attitudes and experiences and the type of explanation that ontogenetics and phylogenetics (learning and natural selection) can provide.

Jacob's book is a densely argued piece of work and the reader is often met with a barrage of aims and claims. It must be stressed, however, that although the full import of Jacob's arguments is not always clear until he comes to sum them up, he consistently argues with philosophical precision and commendable rigour.

Although there is originality in Jacob's book much of it stems from the modification of Dretske's position based on distinctions and arguments made by others in the field. One of the best things about the book is its bringing together in a well organised way the recent large and complicated body of literature on the subject of intentionality and content.

THE UNIVERSITY OF STIRLING FIONA MACPHERSON

Irrational Action: A Philosophical Analysis
By T.E. WILKERSON
Ashgate, 1997. viii + 170 pp. £35.00

There are, apparently, people who, after due deliberation, sincerely believe that on balance they ought to do something—prune their roses, for example (the 'ought' need not be a moral one)—have ample time and the equipment to do so, are neither physically nor psychologically prevented from doing so, yet fail to act. The gardener goes about his other business. His roses go unpruned. How are we to make sense of such a gardener's failure to act? Does the above description of him really make sense? If it does, what stopped him from acting? If it doesn't, where is the offence against logic? In his first chapter, Terence Wilkerson considers Aristotle's answer in Book VII, Chapter iii of the *Nichomachean Ethics*. He also draws on Aristotle's account of the distinction between voluntary and involuntary action in the *Eudemian Ethics*. Aristotle, argues Wilkerson, is right to acknowledge the logical possibility of such backsliding but Aristotle concentrates too narrowly on dramatic cases where someone is so overwhelmed by a very strong emotion that he stops listening to the voice of reason in the heat of the moment and temporarily loses his self-control. In the words of Socrates, he is "hauled about like a slave". Such a person fails to identify his present predicament (an adulterous affair, say) as one proscribed by some general principle to which he consciously subscribes in his calmer moments. His is a story of ignorance of the nature of his present circumstances induced by an overwhelmingly strong emotion. He is like someone drunk or mentally disturbed. But, says Wilkerson, we need to make room for many other cases where, far from being hot and bothered, backsliders are cool, calm and collected, and are fully aware of what they are doing. These cases of backsliding, of 'cool *akrasia*', cannot be accounted for along Aristotelian lines in terms of an internal conflict between the dictates of cool reason and the pull of hot passion because the agents are not experiencing the heat of a hot passion at all. Yet these cool customers fail to do what they know they ought to do. How is this possible?

Wilkerson resists at some length two attempts to belittle the problem of *akrasia*. Chapter 2 argues that, although there is some kind of logical connection between belief and action, it is not as simple as the simple doctrine that sincere belief entails action and therefore the possibility of sincere belief in the absence of action cannot be ruled out on a priori grounds. The penultimate chapter, Chapter 5, resists the view that neither reasons nor actions are commensurable, which, if accepted, would make it logically impossible for a would-be rational agent to weigh the merits of rival calls to action against each other.

Wilkerson distinguishes between several different cases of irrationality. He pictures the idealised rational agent as passing through three stages when deciding what to do. First, there is the deliberative stage in which he has to discover what he really believes and wants and has to weigh their competing claims. Secondly, there is the legislative stage when he has to reach a decision about what to do. In the third stage he has to turn intention into action,

which requires an account from Wilkerson of the role of the will (in Chapter 3). Irrationality arises when the agent fails to negotiate one or other of these three main stages. Failures that occur at the first or second stage are failures of deliberation and form the topic of Chapter 4, 'Self Knowledge'. Failures that occur at the final stage all count as cases of *akrasia*, the subject of the first three chapters.

There are those whose *akrasia* is, as Aristotle said, due to powerful emotions. They wittingly do the wrong thing because they are extremely angry, greedy, jealous, or lustful, etc. Cases of 'cool *akrasia*' arise when one overlooks the importance of one's beliefs which have arisen 'passively' rather than as a result of an active, conscious, choice and one allows them to push one into action; or when one is carried into action by the pull of the present despite one's long-term desires. *Akrasia* due to a defective will arises in those who lack enough will power, or in those who, at the crucial moment, devote all their energies to doing something other than what they know on balance they ought to do.

There are many deft touches of humour throughout this book, including a charming and refreshingly frank "gloomy summary" at the end of the chapter on the will (pp. 89–99). Wilkerson compares himself in his chapter on the will to Harold Macmillan who, as Prime Minister on a visit to the British Embassy in Moscow, which was being bugged by the Russians, was invited to confer with his officials inside a small, collapsible, soundproof wigwam. "And then," Macmillan explained, "no one could think of anything to say." The style is delightful, the mastery of the subject matter obvious yet always understated. The manner is scholarly but there is no hint of scholasticism, either ancient or modern. Only on one occasion did I feel that in his appeals to what it makes sense to say Wilkerson's sureness of touch has deserted him when he argues on p. 11 that there is a sense of 'know' in which it is true that at the time when someone who knows some Russian is concentrating hard on understanding a bank statement, he or she does not know any Russian.

THE UNIVERSITY OF HULL T.S. CHAMPLIN