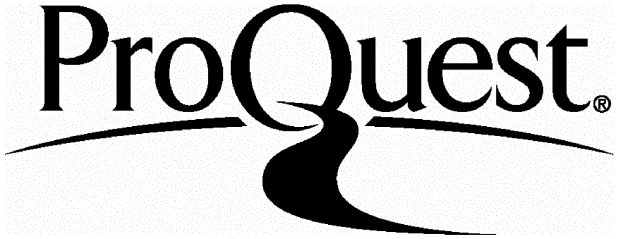
Interpreting Self-Deception

Daniel Reinhold Sascha Viehoff
University College London
M.Phil. in Philosophy



Abstract

My thesis analyses the apparent clash between a central tenet of Davidson's interpretation theory, viz. that our attribution of mental states must make the subject look maximally rational, and our psychological vernacular that allows us to attribute to someone irrational belief states like self-deception.

Self-deception, according to Davidson, involves two contradictory beliefs p and not-p, where not-p comes about as the consequence of an intentional action performed by the agent in order to avoid recognising that p.

This conception of self-deception must explain why holding contradictory beliefs does not lead to the negation of one of them. Davidson addresses this problem by introducing the notion of a 'mental boundary', which holds apart the content of the contradictory beliefs, while allowing their causal interaction. I argue that this reduces self-deception to a 'mental short-circuit', and cannot make sense of our ordinary conception of it. Davidson's reference to intentional action creates a further difficulty: a false belief that has been intentionally brought about lacks a characteristic features of beliefs, their aiming at the truth.

The incapacity to account for self-deception points to a more general difficulty in Davidson's view of the mind. To account for the relationship between first- and third-personal uses of mental terms, Davidson's interpretation theory should be understood in terms of 'imaginative projection'. This highlights a fundamental problem: a theory that takes interpretation to constitute the mental realm cannot explain how the concept of self-deception can come about.

My final chapter analyses the sources of the difficulty: if interpretation aims at formulating 'total theories', it cannot make sense of self-deception, which has local links that are 'rational' in a limited sense but irrational in the bigger picture. To accommodate self-deception, interpretation theory should conceive of interpretation in terms of interest- and occasion-dependent projects.

Table of Content

Introduction

The question that drives this thesis is derived from a puzzle that Davidson poses at the beginning of his 'Deception and Division'[1]. According to him, self-deception and other forms of irrational behaviour set problems for philosophical psychology because

> we find ourselves tempted by opposing thoughts. On the one hand, it is not clear that there is a genuine case of irrationality unless an inconsistency in the thought of the agent can be identified, something that is inconsistent by the standards of the agent himself. On the other hand, when we try to explain in any detail how the agent can have come to be in this state, we find ourselves inventing some form of rationalization that we can attribute to the self-deceiver, thus diluting the imputed inconsistency.[2]

We can frame the problem Davidson thinks he has identified as the difficulty of finding a place for psychologised explanations of behaviour in our conception of human agency. Davidson is not the only philosopher who thinks that instances of irrationality, e.g. self-deception, create conceptual difficulties; it is a common argument against self-deception that the vernacular notion is at best confused, at worst incoherent, and requires serious revision or must be discarded. The most common claim is that it is incoherent to suppose that an agent could simultaneously believe p and not-p, as required on some conceptions of self-deception, or that an agent could form an intention to acquire a belief that she knows to be false, demanded by others. Revisionists among philosophical psychologists conclude that there is no such thing as self-deception; those inclined to protect ordinary ways of talking are more prone to saying that self-deception does exist, but that it is something quite different from what one might think. Lately, the most prominent proposal has been to eschew intentions altogether from an account of self-deception. This 'anti-intentionalist' theory proposes that self-deception must be understood in terms of mechanisms located below the surface level of intentional actions, so that what might seem like the most obvious understanding of the concept – on which deceiving oneself is analogous to intentionally deceiving others[3] – is prohibited.[4]

Davidson sets himself a different task: he has identified a problem with our way of talking about self-deception, but rather than either discard the concept or

---

[1] Davidson (1986); all subsequent references to Davidson's work will use the abbreviations given on p.76.
[2] DD, p.138
[3] See, e.g., Siegler (1963).
[4] The most sophisticated proposal of this kind can be found in Mele (1997, 2001); also see Fingarette (1998). But see Bermúdez (2000) for criticism.

fundamentally change its meaning, he attempts to show that some adjustments to our (philosophical) ways of talking about the mind resolve the issue. Davidson argues that we can alleviate the tension between the rationalising tendencies of our talk about the mental and the existence of irrational behaviour by altering the way we describe mental events.

To heighten our appreciation of the subtleties of Davidson's account, I begin by expounding his picture of the mind, and in particular of rationality, a central notion for Davidson's theory of the mental and his philosophy of psychology. This sets the scene for the second chapter, on Davidson's explanation of self-deception. I discuss how Davidson reshapes our understanding of self-deception, but conclude that his account is ultimately unconvincing because it cannot accurately model the role self-deception plays in our lives.

The third chapter locates Davidson's account of self-deception in the wider context of his account of mental concepts as constituted by interpretation. It highlights that the problem self-deception poses for Davidson's philosophy go beyond the general difficulties identified by other philosophers. I argue that to satisfy certain requirements concerning the relation between first- and third-personal uses of mental terms, we must explicate Davidson's interpretation theory in terms of projective imagination. This reveals that Davidson's account of mental concepts is incompatible with his view on self-deception.

The fourth and last chapter identifies the source of these difficulties: Davidson's understanding of interpretation, according to which interpretation invokes 'global theories', forces on us a picture of the mind as transparent to itself and without space for irrational phenomena. Davidson's attempt to introduce a limited amount of shadow into the translucency of the mind comes too late and is too feeble, and accommodating self-deception in his account of the mind fails. I argue that to make sense of self-deception we must forsake the idea that interpretation invokes 'global theories', and instead conceive of interpretation as a matter of interest- and occasion-dependent projects. This will overcome the problems identified in the previous chapters, while keeping in place the central insights of Davidson's philosophy.

First Chapter: Mind and Rationality

The aim of this chapter is to introduce Davidson's philosophy of psychology and action. This will elucidate Davidson's understanding of the relationship between the physical and the mental; and it will highlight the pivotal role rationality plays in Davidson's picture of the mind.

1. Holism

The thesis of the 'holism' of the mental can be seen as the starting point of Davidson's philosophy.[5] A holistic conception of the mental informs at least implicitly all of Davidson's writing. The incipient idea is that belief and meaning are interdependent notions:

> The interdependence of belief and meaning is evident in this way: a speaker holds a sentence to be true because of what the sentence (in his language) means, and because of what he believes. Knowing that he holds the sentence to be true, and knowing the meaning, we can infer his belief; given enough information about his beliefs, we could perhaps infer the meaning.[6]

The holism that follows from the interdependence of belief and meaning is not peculiar to Davidson, but commonplace in many accounts of the natural and social sciences. Their 'methodological holism' is simply the upshot of the requirement that a whole and diverse body of data should be explained by a unifying theory, and thus a prerequisite of adequate theory construction in these fields. By contrast, Davidson's approach implies that holism is not merely a methodological requirement on any explanation of the mental, but constitutive for what is to be explained: beliefs, desires, intentions etc. are holistically constrained, i.e. they are not independent of the connections which obtain between them. The existence of a mental state depends on other mental states: my belief that I will go for lunch in two hours is connected to my belief that one has lunch around mid-day, or that an hour contains sixty minutes. There is not just an empirical link between these beliefs, i.e. it is not just that people who hold the first one also tend to hold the other ones, but instead the content (and thus the identity) of these beliefs depends on their relationship to each other - mental states can only be understood as part of a network of other such states.

---

[5] Cf. Davidson's remarks about how he came to philosophy from psychology, in, e.g., PP, especially pp.233ff. See also Malpas (1992), p.33 and *passim*; and cf. Ramberg (1989) for an account of Davidson's philosophy of language that gives holism pride of place.

[6] RI, p.135

Davidson's holism is closely dependent on a particular conception of our mental vocabulary. Notions like belief, thought, action, speech, meaning etc. are used to understand and explain human behaviour, including speech behaviour. They take their meaning from the place they have in this project and cannot be understood apart from it:

> We have the idea of belief only from the role of belief in the interpretation of language, for as a private attitude it is not intelligible except as an adjustment to the public norm provided by language. It follows that a creature must be a member of a speech community if it is to have the concept of belief. And given the dependence of other attitudes on belief, we can say more generally that only a creature that can interpret speech can have the concept of a thought.[7]

Similarly, because the identity, i.e. the content and attitude, of a particular mental state (or, as Davidson says, 'mental event') depends on the role it plays in the overall behaviour of the agent, to determine what the mental state is we must also determine what other mental states there are:

> There is no assigning beliefs to a person one by one on the basis of his verbal behaviour, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest.[8]

The interdependence of belief and meaning is only one tenet of Davidson's philosophy; the other is the pairing of meaning and truth.[9] Interpretation of behaviour depends on the notion of truth, because it requires stating when someone else's utterance would be true, where the conditions for the truth of the statement are expressed in a language that the interpreter already understands. For Davidson, truth-conditions are assigned by pairing utterances with situations or circumstances in which they are used, where the pairing is constrained by a demand for overall coherence and consistence. An important consequence of this account of assigning truth-conditions is the presumption that the speaker gets things by and large right: if the interpreter uses relations between the speaker's utterances and the environment to determine the appropriate interpretation of the speech act, she must assume that the speaker frequently speaks the truth about the matter.

Finally, the notions of belief and truth are closely related for Davidson: we need a concept like belief because we interact with others and thus experience intersubjectivity. This experience requires that there could be a gap between different people's beliefs about the world, and thus also between beliefs and what

---

[7] TT, p.170
[8] ME, p.221
[9] Since this is an essay about Davidson's philosophy of psychology, I do not try to defend Davidson's semantic theory, and restrict myself to outlining those bare essentials that are important for understanding Davidson's philosophy of mind.

they are about. The notion of truth in turn rests on intersubjective agreement and could not exist without it.[10]

Together, the interdependence of belief, meaning, and truth yields a conception of understanding – language, others, and the world – as arising from our simultaneous interpretation of the linguistic and non-linguistic behaviour of others, their attitudes, and the environment.

2. Interpretation

For Davidson, propositional attitudes, or intentional states, are central for the realm of the mental; only someone who possesses propositional attitudes counts as a thinking being, a mind in the proper sense of the word.[11] It is unclear whether Davidson thinks that sensations like pain, hunger, or moods, which are not propositional in character, are not 'properly speaking' part of the mental; sometimes his writing seems to suggest as such. But this conclusion is not *per se* entailed by his theory; we could also interpret his insistence on the priority of propositional attitudes as a rhetorical strategy, a matter of drawing attention to those areas where what he views as characteristic of the mental is most obvious. This would not undermine the possibility that moods and pains are also properly mental[12], but it would leave them on the argumentative backseat because there is the prejudice that they are not covered by what Davidson thinks is the defining characteristic of the mental: rationality. Rationality determines the relations between content-bearing states, so the initial focus on propositional attitudes brings to light its centrality for our picture of the mind.

The normativity of the mental plays out in our interpretative practices: according to Davidson, a rational reconstruction of our practice of attributing psychological states necessitates the interpretation of other people's behaviour, where interpretation is a matter of redescribing that behaviour in terms that are 'revealing' and comprehensible. The point of interpretation, and of our psychological vocabulary employed in interpretation, is to 'make sense' of the behaviour of others, which is a matter of fitting the behaviour of others into a pattern of

[10] Cf. RA, especially pp.104f.
[11] Cf. RA, *passim*
[12] If we accept that there is more to rationality than the requirements of deductive rationality to which we can give full linguistic expression, we can make sense of the idea that emotions and moods can 'fit the occasion' and thus be 'reasonable'. Such normative perspectives on non-propositional attitudes could also circumscribe their attribution.

explanation constrained by the requirement that their attitudes, beliefs, thoughts etc. be coherent and consistent.

For Davidson, rationality is not limited to the relations between beliefs alone, but covers all behaviour, including, e.g., links between beliefs and actions. As the cogency of our ordinary explanation of human behaviour rests on the possibility of discovering coherent and consistent patterns in the agent's linguistic and non-linguistic behaviour, whenever we deal with other people as psychological rather than purely physical beings the requirements of rationality apply:

> Coherence here includes the idea of rationality both in the sense that the action to be explained must be reasonable in the light of the assigned desires and beliefs, but also in the sense that the assigned desires and beliefs must fit with one another.[13]

3. Reasons and Actions

The most explicit account of the central role rationality plays in understanding others is provided by Davidson's theory of reasons and actions. For a belief to be the reason for an action two conditions must be fulfilled: (i) the link between the belief and the action must be 'rationalisable', that is, we must be able to describe it with a practical syllogism involving a desire and a belief which together explain why the agent undertook the action. The practical syllogism is subject to the requirements of rationality – "In the light of a primary reason, an action is revealed as coherent with certain traits, long- and short-termed, characteristic or not, of the agent, and the agent is shown in his role of Rational Animal"[14] – and it must be able to capture the agent's perspective: "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in the action."[15] Yet although these conditions pick out 'candidates for reasons', they cannot distinguish between different candidates and determine which one of them is the 'real' reason. Davidson therefore imposes an additional condition: (ii) the 'reason for action' is distinguished by its singular causal role, only *this* reason caused the action.

Davidson holds that causal explanations must be backed up by nomological statements, which apply to whole classes of events. The underlying idea is that causal connections exist between events however they are described, whereas causal laws, and thus explanations, apply to events under specific descriptions. Since such causal laws are not available in psychology, Davidson must invoke physical terms in talking about actions and reasons:

---

[13] TT, p.159
[14] ARC, p.8

> The law whose existence is required if reasons are causes of actions do not, we may be sure, deal in the concepts in which rationalization must deal. If the causes of a class of events (actions) fall in a certain class (reasons) and there is a law to back each singular causal statement, it does not follow that there is any law connecting events classified as reasons with events classified as actions – the classifications may even be neurological, chemical, or physical.[16]

Our ignorance of laws formulated in neurological, chemical, or physical terms does not inhibit our ability to provide valid causal explanations, and thus cannot count as an argument against conceiving of reasons as causes. The validity of most causal explanations, Davidson holds, depends not on covering laws that are fully spelled out, but on evidence that points towards the existence of such laws and their applicability to the case at hand. While it is true that "'A caused B' entails that there exists a causal law instantiated by some true descriptions of A and B"[17], it is also true that "no particular law is entailed by a singular causal claim, and a singular causal claim can be defended, if it needs defence, without defending any law."[18]

4. Mental Events

Davidson's account of explanations of actions adumbrates his overall position with respect to the mind. The commitment of the mental to the notion of rationality, and the consequent explanatory practice that employs terms like 'reasons' and 'beliefs', makes a reduction of the mental to the physical impossible. But Davidson does not defend a dualism of the mental and the physical; instead, he argues for a position that is physicalist and leaves room for the causal interaction of mental and physical events, yet excludes the possibility of psychophysical laws. His 'anomalous monism' states that mental events are physical events – hence Davidson is a monist rather than a dualist – but also argues that mental events cannot be explained by physical events.

Events are "unrepeatable, dated individuals"[19], and they are mental if their description essentially contains at least one mental verb – that creates an apparently non-extensional context, i.e. gives rise to the problem of substitutability – and physical if they are picked out by sentences containing only physical vocabulary.[20] The same event can thus be mental (when described in a mental vocabulary) and

---

[15] Ibid., p.3
[16] Ibid., p.17
[17] Ibid., p.16
[18] Ibid., p.17
[19] ME, p.209
[20] Davidson recognises that with appropriate descriptions every event can be counted as 'mental', but does not consider this a difficulty for his account: if he can show that despite the excessive range of mental events his theory holds, he can claim greater strength for it.

physical (when described in a physical vocabulary), and monism follows from Davidson's assumption that all events are physical. Causal relations exist between events no matter how these events are described, so the distinction between the mental and the physical, which on this account is merely a matter of description, cannot affect the causal efficacy of events. By contrast, the linguistic character of nomological explanations, i.e. the fact that explaining something requires language, makes their adequacy conditional on how the events in question are described. Davidson thinks of explanation as the explication of causal relations between events; because his conception of causality requires that every causal statement could be backed up by a strict nomological statement, successful explanation depends on the existence of strict causal generalisations between the types of events in question. The 'anomalousness' of the mental can thus be framed in terms of 'types of events': although mental events are identical with physical events, no type of mental event is identical with a type of physical event. By implication, although mental events are identical with physical events, and mental and physical events interact causally, there are no psychophysical laws, because laws are linguistic in nature and cover relations between types of rather than individual events.

The dichotomy of the mental and the physical is thus a dichotomy of vocabularies: we cannot explain or predict mental events in terms of physical events, or at least not offer the kind of causal laws that Davidson thinks are required for such an explanation to be successful, because "Nomological statements bring together predicates that we know a priori are made for each other... [and] mental and physical predicates are not made for one another."[21]

One particularly important – and contested – conclusion that one must draw from Davidson's characterisation of the respective domains of the mental and the physical concerns the possibility of explanation: since explanations are framed in language, the occurrence of an event can only be explained if the event is given under a particular – mental or physical – description. For Davidson, the paradigm for any explanation is the explanation of physical events in terms of their causal connections, and this requires citing causal laws of which the case in question is a particular instance. Since only the causal relations between physical events satisfy the stringent conditions for properly nomological explanation, the only truly satisfactory explanation refers to physical events.

[21] Ibid., p.218

Davidson qualifies this claim though: he allows that certain types of explanations can be cast in mental terms, and that "if an event of a certain mental sort has usually been accompanied by an event of a certain physical sort, this often is a good reason to expect other cases to follow suit roughly in proportion."[22] These explanations rely on generalisations which do not satisfy the stringent conditions of counterfactuality Davidson introduces for nomological statements. Thus their role is not to provide a truly sufficient explanation, but to lend support to singular causal claims and related explanations of particular events. "The support derives from the fact that such generalizations, however crude and vague, may provide good reason to believe that underlying the particular case there is a regularity that could be formulated sharply and without caveat."[23] The rough generalisations nonetheless provide the basis for our ordinary explanation of mental events:

> ...it is possible ... to know of the singular causal relation without knowing the law or the relevant descriptions. Knowledge requires reasons, but these are available in the form of rough heteronomic generalizations, which are lawlike in that instances make it reasonable to expect other instances to follow suit without being lawlike in the sense of being indefinitely refinable.[24]

There could not be stronger laws because the domain of the mental is not the 'comprehensive closed theory' that Davidson argues is required for strict causal laws to hold. Causal explanations of mental events in terms of physical events are also impossible because explanations in these two domains are governed by different principles:

> There are no strict psychophysical laws because of the disparate commitments of the mental and the physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual. There cannot be tight connections between the realms if each is to retain allegiance to its proper source of evidence.[25]

5. The Irreducibility of Rationality

Davidson suggests that the commitment of the mental to rationality makes it impossible to reduce it to or explain it in terms of physics. Brian Loar (1981) argues that this conclusion is mistaken: "Can a functionalist theory, that each attitude-sentence ascribes a second-order property expressible in a physical vocabulary, recognize the constitutive force of rationality? Of course it can, and must."[26] This conclusion follows, I argue, from a misunderstanding of Davidson's theory and the

---

[22] Ibid., p.219
[23] Ibid., p.219
[24] Ibid., p.224
[25] Ibid., p.222
[26] Loar (1981), p.22

role rationality plays in it; identifying where Loar's argument goes wrong will help clarify the assumptions underlying Davidson's position.

Loar aims to show that a functionalist theory, proceeding holistically in the ascription of mental roles to the physical base, could incorporate counterfactuals mirroring the restrictions imposed by rationality on beliefs and other attitudes: "...a system of physical state-types satisfies the constraints on rationality provided they are all related counterfactually as the theory says the beliefs to which they correspond ought rationally to be related."[27] The counterfactuals would restrict the possibility that particular physical states occur, and thus reflect the structure imposed upon beliefs – that is, for Loar, physical states described according to their functional role – by the norms of rationality. Obviously, a description cannot affect the existence or non-existence of a physical states; whether a certain relation between physical states exists is only contingently and not necessarily true, yet the relation between beliefs that are rationally linked is by no means contingent. Loar bridges the gap between the contingency of physical states and the non-contingent character of mental states by making contingent the identification of a particular physical state as a mental state:

> It is a priori that if certain states are to be counted as beliefs and desires they must satisfy the constraints on rationality. But that they do satisfy them can be as contingent as you like; if they fail to do so they are not beliefs and desires. It is a fallacy to argue that, since rationality has constitutive force, a physical system would have to conform to rationality non-contingently for its workings to constitute the workings of a mind.[28]

Thus, Loar grants that rationality has a constitutive role of some sort for our talk about propositional attitudes, but for him this by no means implies that the mental is irreducible to the physical. His argument rests on the formulation of meaning postulates (or 'M-constraints') and certain basic constraints on the relations between beliefs – specifiable in terms of their logical form – which he calls 'L-constraints', and which for him are sufficiently close to (deductive) rationality to model our mind. Yet that these L-constraints cannot capture the full structure of deductive rationality is easily shown: deductive rationality determines what is logically valid, and there is no mechanical proof for logical validity in general; but if the whole of deductive rationality could be spelled out in terms of rules like Loar's L-constraints, such a mechanical proof would be available; hence L-constraints can never fully capture the structure of deductive rationality. Loar does not deny this; but he

---

[27] Ibid., p.22
[28] Ibid., pp.23/4

believes that the thinned-down version of deductive rationality suffices to make sense of our mental lives.

The divergence between Loar's conclusion – that functionalist theories of the mind can succeed irrespective of the constitutive role of rationality for the mental – and Davidson's view that the constitutive role of rationality precludes any reduction of the mental can be traced to differences between Loar's and Davidson's understanding of rationality.[29] The basic constraints on the relations between beliefs that Loar offers are intended to reflect 'deductive rationality', i.e. the capacity to hold beliefs because they follow deductively from other belief one already holds. That L-constraints cannot capture the full structure of deductive rationality may initially not disqualify them from elucidating the mental realm: deductive rationality is only imperfectly instantiated in human beings, and Loar's conclusion that the constitutive force of rationality for the realm of the mental depends on some minimal requirements on the structure of what can be recognised as a mind might therefore seem plausible. We could assure ourselves of the reality of rationality by ascertaining the satisfaction of particular structural conditions by what counts as a rational mind. Even Davidson's statement that "when we use the concepts of belief, desire, and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory"[30], could be accepted on the premise that what is meant by 'rationality' is nothing but Loar's set of structural constraints on what counts as beliefs. If 'rationality' is read in this fashion throughout, then it seems possible to explicate the mental in functionalist, and eventually in purely physical, terms.

But in spite of this initial success, Loar's restricted understanding of rationality cannot make sense of certain important features of our mental lives. Outside of the sphere of beliefs related by L-constraints, Loar's theory could not deliver an understanding of beliefs as dependent on other beliefs that are deductively cogent reasons for them. Furthermore, because the constraints cannot reflect the full structure of deductive rationality, Loar's theory lacks the general normative notion of deductive consequence, and the explanations it can offer must instead rely on certain transitions that the mind is, as a matter of fact, prone to making, rather than transitions which are rationally required yet not captured in L-constraints.

---

[29] My argument here follows closely McDowell (1985)'s critique of Loar.

It may seem as if this is only a minor quibble, which Loar could overcome by adding further L-constraints. But this response ignores a more fundamental problem: Loar's account cannot accommodate the inherently normative character of rationality and beliefs, because it cannot explain in which sense, if any, beliefs 'ought' to be related as stipulated by the constraints Loar offers. If (a part of) the notion of rationality is taken to depend on the existence of these constraints, as Loar implies, then nothing over and above these constraints can determine our understanding of (that part of) rationality; in particular, we would be unable to criticise the beliefs that are constrained in this fashion, because there is no further standard of rationality we could invoke. If rationality were conceived in terms of Loar's functional constraints, we could not explain a situation in which someone's behaviour, though in line with L-constraints, would fall short of our ordinary notion of rationality and leave us without proper understanding of what is going on.

If the functionalist constraints cannot be criticised, they cannot be justified either: to the extent that for Loar the rational features of the mind are the upshot of, and nothing over and above, these constraints, the form they take – and thus the relation between beliefs that follows – cannot be justified in the way we would expect, viz. with reference to rationality and good reasoning. Thus, Loar's argument for the role of rationality – that we would not recognise something that fails to accord with these constraints as a mind – must be spelled out in counterintuitive terms: for Loar, it just so happens that we cannot recognise something as a mind if it does not satisfy the L-constraints. Rather than say that the constraints he proposes are important because they reflect what is rational, Loar's suggestion seems to be that rationality is important because our mind is constrained in this particular fashion. He thus foregoes the more obvious and satisfactory explanation: that we ultimately justify our failure to understand someone by saying that what she says or does lies "outside the boundaries of what is intelligible"[31]. This straightforward claim requires that we conceive of rationality as an inherently normative notion that transcends our ability to recognise certain things as minds.

If we extend our view beyond the realm of deductive reason, it becomes even more obvious that rationality, understood as the constitutive force behind the mental, must be 'transcendent' and ultimately inexplicable in non-normative or non-intentional terms. Two important phenomena cannot be explained if we

[30] ME, p.223

assimilate explanation that depends on the ideal of rationality to that which subsumes events under patterns of what generally tends to happen. First, we are robbed of the critical dimension inherent in our ordinary explanation of human actions. Reasons people have for believing or acting are good or bad reasons, yet there is no way of specifying what 'good' and 'bad' reasons are without reference to intentional or normative notions; hence a functionalist account of the mind, which tries to explicate Davidson's rationality requirements in terms of functional constraints, cannot elucidate our talk of 'reason'. Second, the perspective of the acting and deliberating subject can only be properly understood if we take her to adhere to the ideal of rationality: "Without any external touchstone, there seems to be no ground on which a subject or group could be confident that its own grasp of the structure, from inside, was incapable of improvement, in particular from coming to understand others."[32] In other words: if there were no such 'transcendent' notion of rationality, the very idea of improving upon one's previous reasoning or deliberation would be unexplainable, because there would not be any standard against which such an 'improvement' could be measured.

Rather than a minimal requirement on what counts as a mind, formulated in terms of basic rules of deductive rationality instantiated in the behaviour of what we happen to recognise as minds, Davidson requires the full structure of rationality – deductive and other – to serve as a constitutive ideal for the mental realm, because, as McDowell explains, propositional attitudes have their place in "explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be."[33] An understanding of rationality that ignores this transcendent character assimilates this special kind of explanation with another kind, which makes things intelligible "by representing their coming into being as a particular instance of how things generally tend to happen"[34]. Loar's L-constraints fall short of rationality in the 'transcendent' sense – i.e. as never fully explicable in terms of non-intentionally specified rules – that Davidson's theory relies on, and we must conclude that the revisionism that informs Loar's arguments must be alien to a philosopher like Davidson who holds that "nomological slack between the mental and the physical is essential as long as we conceive of man as a rational animal."[35]

---

[31] McDowell (1985), pp.330/1
[32] Ibid., pp.331/2
[33] Ibid., p.328
[34] Ibid., p. 328
[35] Ibid., p.223

6. Conclusion – Persons and Rationality

In section 2 we saw that Davidson thinks of the mental as unified by the commitments embodied in the language used to talk about it. 'Mental' events count as such because they are described in a fashion that makes essential use of terms that give rise to problems of substitutability; so the concept of the mental depends on the vocabulary that is being used to talk about mental events. Davidson's 'monism' furthermore entails that this vocabulary is not the only one available to talk about the events in question: they are simultaneously mental and physical, because both vocabularies can be used to describe them. The anomalousness of the mental precludes that we bit by bit substitute physical terms for our ordinary mental vocabulary – along the lines envisaged by Stich and Sellars[36]. But this does not yet provide us with a fully fleshed out explanation of why, according to Davidson, we use the mental vocabulary in the first place.

A first hint, that will lead us into our discussion of irrationality in the next chapter, is offered by Davidson's claim that the irreducibility of theories of belief and meaning depends on the methods we must invoke in constructing them[37]:

> Each interpretation and attribution of attitude is a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth, and it is this that sets these theories forever apart from those that describe mindless objects, or describe objects as mindless.[38]

So the commitment to rationality and holism – which are really two sides of the same coin – lie at the base of our use of mental terms.

One important insight we should take away from this is that attributing rationality to someone is not a psychological matter, but a logical one – not a matter of being nice, but of having towards them the attitudes we characteristically have only towards person:

> Crediting people with a large degree of consistency cannot be counted mere charity: it is unavoidable if we are to be in a position to accuse them meaningfully of error and some degree of irrationality. ... To appreciate the limits of the kind and amount of blunder and bad thinking we can intelligibly pin on others is to see once more the inseparability of the question what concepts a person commands and the question what he does with those concepts in the way of belief, desire, and intention. To the extent that we fail to discover a coherent and plausible pattern in the attitudes and actions of others we simply forego the chance of treating them as persons.[39]

---

[36] Stich (1983), Sellars (1997), especially pp.80-115
[37] BBM, p.154
[38] Ibid., p.154
[39] ME, pp.221/2

Treating people as persons requires imputing rationality to their words and deeds. The notions of personhood and rationality, and the possibility of understanding other people as persons whose behaviour bears meaning, i.e. of interpreting them and identifying meaningful patterns that warrant attributing attitudes and content to others, are therefore all closely interlinked. Persons, but not things, are subject to criticism for having the wrong reasons, or being irrational, so that our reliance on the notion of personhood is of a piece with the idea that there is something that is governed by laws of reason and rationality. So the ultimate reason for using notions like 'belief', 'desire', 'meaning', 'hope', 'thought' etc. is our way of looking at the world as something in which notions like rationality – norms – have a place.

Second Chapter: Self-Deception *en detail*

In the last chapter I expounded Davidson's view on the mental and analysed the particular role of rationality in his accounts. In this chapter we turn to a phenomenon that seems to pose a serious problem for theories that put as much of a premium on rationality as Davidson's: irrational behaviour. Our analysis focuses on Davidson's account of self-deception – frequently labelled the 'standard approach'[40] – as developed in his 'Deception and Division'. I argue that the standard approach contains an irresolvable tension: to explain self-deception, Davidson introduces the notion of a mental boundary across which only causal, but not intentional or rational, links are possible; but this mental boundary prevents both intentions and the contradictory nature of the beliefs involved in self-deception from playing the role that Davidson assigns to them.

1. Davidson's Model

At the inception of 'Deception and Division' lies a difficulty in explaining self-deception and other forms of irrational behaviour: we are relatively confident in attributing self-deception to an agent, but when we try to explain how the agent can be in this state, we find ourselves rationalising her behaviour and stripping it of its irrational character. Davidson's task is to provide a model of self-deception that is both explanatorily satisfactory and does not expunge the irrationality in question. Davidson's account is ingenious and subtle; he puts the irrationality of the phenomenon at centre stage by describing it as a contradictory belief state, and sets it apart from other such states by an intricate account of its aetiology.

In the (particularly troublesome) cases of self-deception Davidson focuses on, an agent holds a belief, or believes that she holds sufficient evidence to warrant a belief, p, and this belief (or what the agent thinks is evidence for it) causes her to hold a contradictory belief. She thus simultaneously holds two contradictory beliefs[41], one of comes about in response to the other. Davidson takes it for granted that it is impossible to hold the belief 'p and not-p', so the beliefs must somehow be kept apart to prevent the agent from noticing the contradiction. We can distinguish three steps in the process of belief formation:

---

[40] Cf. Dupuy (ed.) (1998), p.x and *passim*

[41] To avoid prolixity, I assume that talk of 'contradictory beliefs' covers both explicitly contradictory propositions (such as p and not-p) and a belief and beliefs that are evidence for its negation.

First step: the agent "A has evidence on the basis of which he believes that p is more apt to be true than its negation".[42] This belief motivates the second step.

Second step: the agent A acts "in such a way as to cause himself to believe the negation of p."[43] The act is motivated "by the thought that p, or the thought that he ought to act in such a way as to cause himself to believe the negation of p", and performed "with the intention of producing a belief in the negation of p."[44] "The action involved may be no more than an intentional directing of attention away from the evidence in favor of p; or it may involve the active search for evidence against p."[45]

Third step: the agent believes the negation of p. Yet he also believes p, or that he has stronger evidence for p than for not-p, since the act that brought about and keeps in place the belief that not-p depends on his believing p. This final state is troublesome because "the self-deceiver knows he has better reasons for accepting the negation of the proposition he accepts...: he realizes that conditional on certain other things he knows or accepts as evidence, the negation is more likely to be true than the proposition he accepts; yet on the basis of a part only of what he takes to be the relevant evidence he accepts the proposition."[46] Davidson calls this failure of reasoning 'weakness of the warrant'; it amounts to sinning against the 'requirement of total evidence for inductive reasoning', i.e. the principle that in deciding between a set of mutually exclusive hypotheses, we ought to give credence to the hypothesis most highly supported by the overall available evidence. So for Davidson, self-deception is usually an instance of weakness of the warrant that is self-induced or motivated; these are the cases we will concentrate on in the following discussion.

2. Contradictions and the Mental Boundary

Davidson anticipates resistance to his description of the third step, which seems to involve blatantly contradictory, and hence self-defeating, beliefs: as Davidson acknowledges, we cannot attribute to someone the belief 'p and not-p' even if she utters these very words. It is out of line with the most basic requirements of rationality to hold obviously contradictory beliefs, and an interpretation of another's actions and words that would yield such patent violations of rationality would be

[42] DD, p.145
[43] Ibid., p.145
[44] Ibid., p.145
[45] Ibid., p.145
[46] Ibid., p.142

found wanting. Davidson insists nonetheless that the agent who deceives herself holds such contradictory beliefs, and sets out to show that it is possible to attribute them to her in a way that does not yield a flawed interpretation.

At this point it is important to draw a distinction between two different understandings of 'contradictory beliefs': it seems clear that the belief 'p and not-p' is contradictory, and that rationality excludes this belief from being attributed to someone whose behaviour we interpret. But it seems equally obvious that people hold (probably a whole range of) beliefs which, when spelled out properly, are contradictory, but which have not yet been spelled out adequately and may never be. Otherwise we would have to assume that people can check each new belief against every other belief they already hold, and economy and experience alike suggest that this does not usually occur. What sets self-deception apart from these more common cases is that in self-deception the contradictory beliefs are not acquired independently of each other; instead, they are linked by an intention, yet their coexistence does not lead to the cancelling out of one by the other. In addition to this analytical distinction, we can also draw a 'phenomenological' distinction between the two kinds of contradictory beliefs: when agents hold the common, quite innocuous kind of contradictory beliefs, simply pointing out to them that their beliefs are contradictory will usually induce them to give up one of them; if, by contrast, the agent holds contradictory beliefs as a result of self-deception, pointing out the contradiction will not do the job, as the agent will defend her position and refuse to acknowledge her mistake.[47] Since only the latter kind of contradictory beliefs matter for Davidson's argument, for the remainder of this thesis I will refer to these when I mention 'contradictory beliefs'.

Both Davidson's argument and my discussion focus exclusively on beliefs one could call 'avowable'[48]: the agent can become non-inferentially aware that she holds this belief. In certain circumstances, beliefs cannot be accessed in this fashion; for example, in therapy the agent only recognises that she holds certain beliefs on the basis of evidence that the analyst points out to her. In such cases, her knowledge of the belief in question is broadly similar to the knowledge she has of other people's beliefs: she draws on all the available evidence, and concludes that the best understanding of her own behaviour invokes a belief of which she had not become aware through normal channels. Beliefs that are attributed to oneself on the

[47] This observation is discussed in more detail in section 5.

basis of evidence, i.e. because they provide the best explanation of one's own behaviour, may be important for overcoming self-deception; but they are irrelevant for Davidson's account of the phenomenon as involving contradictory beliefs which the agent could not usually hold simultaneously. Beliefs attributed to an agent on the basis of evidence and inaccessible through the ordinary channels cannot lead to the kind of contradiction between beliefs that is required for Davidson's account of self-deception. In the remainder of our discussion, 'belief' thus refers to ordinary, first-personally accessible beliefs.

Davidson's strategy to solve the puzzle of self-deception depends on the distinction he draws between the logical (or rational) and the causal force of reasons. As discussed in the previous chapter, for Davidson some of the qualities of events are independent of any description, e.g. their causal force, while other qualities differ according to how the event is described, e.g. as mental or physical. Since events can be described in different ways, they can be simultaneously mental and physical.

The upshot of Davidson's doctrine is that beliefs, reasons, desires etc. have both rational and causal force. If we identify a contradiction between two beliefs, then we must focus on their rational force because events described in physical terms are not subject to rational scrutiny, and hence cannot be logically contradictory. When belief p motivates an action that brings about belief not-p, the link between the two must be causal, but cannot be rational, since we cannot envisage that holding belief p *rationally* warrants anything that leads to holding the belief not-p.

Davidson therefore wants to "find a point in the sequence of mental states where there is a cause that is not a reason; a specific irrationality by the agent's own standards of rationality."[49] He identifies this specific irrationality with the event that makes it possible to hold two contradictory beliefs: the emergence of a 'mental boundary' that keeps the two inconsistent beliefs apart. Davidson can postulate the existence of mental boundaries because they are "conceptual aids to the coherent description of genuine irrationalities"[50] rather than introspectively ascertainable mental phenomena.

---

[48] Cf. Moran (2001), pp.83ff., for the relevant distinction between avowals and reports.
[49] DD, p.145
[50] Ibid., p.147

Yet the precise function of the mental boundary is difficult to assess as Davidson offers two different, and not obviously equivalent, explanations. One characterisation runs like this:

> How can a person fail to put the inconsistent or incompatible beliefs together?
> It would be a mistake for me to try to answer this question in a psychologically detailed way. The point is that people can and do sometimes keep closely related but opposed beliefs apart. To this extent we must accept the idea that there can be boundaries between parts of the mind; I postulate such a boundary somewhere between any (obviously) conflicting beliefs.[51]

Note that the boundary is drawn between contradictory beliefs in this case. Yet at the end of his essay, Davidson concludes that

> In the case where self-deception consists in self-induced weakness of the warrant what must be walled off from the rest of the mind is the requirement of total evidence. What causes it to be thus temporarily exiled or isolated is, of course, the desire to avoid accepting what the requirement counsels. ... In the extreme case, when the motive for self-deception springs from a belief that directly contradicts the belief that is induced, the original and motivating belief must be placed out of bounds along with the requirement of total evidence.[52]

This characterisation is not obviously compatible with the previous one, since Davidson explains that 'having a principle' like the requirement of total evidence does not require that the agent articulate or be aware of it, but only that her "pattern of thoughts [is] in accord with the principle"[53]. If this suffices, then it is dubious whether the principle in question is to count as a belief.

More importantly, it is not evident that the requirement of total evidence could contradict another belief in the sense required here, as it does not apply to individual beliefs but instead governs the relationship of evidence to conclusion. Once the principle of total evidence is reified and taken as something that can be 'exiled', something that can be on one side of the mental boundary with beliefs on the other side, we have to model the relations between them in terms of a triangle of evidence, conclusion, and the principle of total evidence. Only if we have knowledge of all three points can we determine whether there is a contradiction between the mental states of the agent so that consistently describing them requires the concept of a mental boundary. There should therefore be no need to sequester more than one of the three to avoid inconsistencies: if the evidence is 'out of reach' of the requirement of total evidence and the conclusion, then no contradiction can obtain, and if the requirement of total evidence is sequestered, there is no need to locate a boundary between evidence and conclusion.

---

[51] Ibid., p.147
[52] Ibid., p.148
[53] Ibid., p.141

To avoid this difficulty, a satisfactory interpretation of Davidson's remarks about the sequestering of the requirement of total evidence should emphasise its 'relational' nature: the only way to find out whether an agent in general or in particular cases satisfies the principle is by attending to the available evidence and the conclusion she draws from it. Consider again the function of the principle: "...when we are deciding among a set of mutually exclusive hypotheses, this requirement enjoins us to give credence to the hypothesis most highly supported by all available evidence."[54] If the principle is not at work, then, Davidson seems to say, the conclusion drawn is not the one most highly supported by all the available evidence. This interpretation captures the normative nature of the requirement of total evidence, viz. that the agent should consider all the evidence she could potentially use, and makes clear that the agent violates a basic requirement of rationality. If we accept this, then we should admit no distinction between 'exiling evidence' and 'exiling the principle', because the criterion for 'exiling evidence' is that conclusion and available evidence do not square, and this *prima facie* amounts to a violation of the requirement of total evidence. So the requirement can also be said to be 'exiled behind the mental boundary', and the ambiguities in the characterisations of the mental boundary can be straightened out.

According to Davidson, mental boundaries make it possible to consistently and non-contradictorily talk about self-deception without making it more rational than it is, because talk of mental boundaries allows us to distinguish between the causal and the rational force of mental states: while the evidence for p does not serve as a reason for believing that p, it causes the formation of the contradictory belief not-p. In more extreme cases of self-deception, the belief p itself causes the formation of the belief not-p; not-p is not discarded because the beliefs are insulated by a mental boundary. Beliefs, including contradictory beliefs, can thus be causally related without being rationally linked – belief p can cause belief q without making it reasonable to hold q. The upshot of this thesis is that we should not expect causal relations to be isomorphic with rational relations[55]; in particular, the causal relation between noting evidence and forming a belief will not always map neatly onto the logical or rational relations between good evidence and the conclusion this evidence warrants.

[54] Ibid., p.140

3. Belief and Evidence

Davidson's account of the mental boundary entails a particular conception of evidence and its link to beliefs. First, for Davidson, only beliefs held by the agent and taken by the agent to support or disprove a particular hypothesis are evidence for that hypothesis; beliefs which others know warrant the conclusion, but are not held by the agent, do not count as evidence in this sense.[56] Second, it is possible for the agent to take something to be evidence for a certain conclusion, yet not draw that conclusion from it. In the last section we saw that for Davidson, someone who deceives herself draws a conclusion that disregards the requirement of total evidence, i.e. ignores either some or all of the evidence. (The latter option leaves it open whether the agent somehow draws the appropriate conclusion, but then ignores or drops it at the expense of the false one.)

However, when we combine Davidson's conception of evidence, interpretation theory, and the possibility that the agent fails to apply the requirement of total evidence, we encounter a problem. According to interpretation theory, the content of propositional attitudes attributed to an agent depends on their place in a network of other propositional attitudes and the world. For something to count as evidence in Davidson's sense, the agent must believe that there is a certain state of affairs *and* that this state of affairs provides rational support for the conclusion that a certain belief p is true. The difficulty is to see how we can justify attributing the second conjunct if the agent does not use the belief as evidence for the conclusion. This problem is compounded when, as in Davidson's account of self-deception, not only does the agent not draw the conclusion p from what she takes to be evidence for p; she actually draws the contradictory conclusion not-p. It appears paradoxical that evidence for a proposition p sometimes causes the belief p and sometimes the belief not-p. If the evidence acquired *warrants* a belief, then we would expect that under normal circumstances the agent also forms that belief rather than another one that is not warranted. We will return to this problem later in this thesis; for now we will simply note the difficulty it poses for Davidson.

If we think of self-deceptive belief formation as a matter of one belief causing, but not rationalising, a contradictory belief, then we must assume that these beliefs would ordinarily be incompatible. The attribution of the belief 'this is evidence for p' to the agent who avows 'not-p' rests on the assumption that in ordinary

---

[55] Cf. Belgum (1990), pp.134/5

circumstances, i.e. were she not deceiving herself, the agent would draw the conclusion p from the facts presented to her. It excludes other options for interpretation: the agent does not simply misjudge the evidence, the juxtaposition of evidence for p and the formation of the belief not-p is not a mere coincidence, nor is it just another example of the ever present possibility of holding contradictory beliefs in the innocuous sense discussed earlier. Thus, Davidson's account of self-deception requires not merely the co-presence of contradictory beliefs, but also a link between them. This only seems convincing if p and not-p are connected not only in spite of, but due to their contradictory contents.

The link between them cannot be an expression of ordinary rationality; nor can it be purely causal or physical: the feature that distinguishes self-deception from other forms of irrationality is precisely that a belief arises as a result of another belief it contradicts. If self-deception is explained in physical terms alone, it becomes a mere coincidence that a particular instance of irrationality also happens to be a case of self-deception, because the causal links are, on this picture, fully independent of the rational links, i.e. the fact that the beliefs are contradictory.

Against this background, we can make sense of Davidson's insistence that self-deception must be intentional: if there were no intentions involved, the evidence for p could be assumed to cause the belief not-p in the same way in which, on other occasions, it causes the belief p, and this would undermine the interpretation of them as 'evidence for p' and 'not-p' on ordinary standards of interpretation. So intentions need to be introduced to ensure that the link between the evidence and the belief acquired is less direct, and hence the attribution of these beliefs to the agent more plausible.

4. The Intentional Action

The second step in Davidson's account requires that the agent acts so as to cause herself to acquire the belief not-p, where her action is motivated by her initial belief p (or strong evidence for p). "The action involved may be no more than an intentional directing of attention away from the evidence in favor of p; or it may involve the active search for evidence against p. All that self-deception demands of

[56] Cf. DD, p.140

the action is that the motive originate in a belief that p is true ... and that the action be done with the intention of producing a belief in the negation of p."[57]

Davidson does not explicitly justify the introduction of an intentional action into the model; in addition to the potential solution intentions offer for the problem discussed in the last section, they may for Davidson simply be required by our ordinary conception of self-deception. This view might seem reasonable if we take into account that the agent who deceives herself is subject to severe (moral) reproach[58] that goes beyond what appears justified by lazy thinking and could not be squared with the idea that in self-deception the agent is victimised by some external force.[59] If, as it often seems in discussions of self-deception, the only alternative to introducing intentions is to explain self-deception in terms of a sub-personal mechanism that the agent has no power to control, and for which she is not responsible, intentionalism may be required to make sense of the fact that self-deception has a moral dimension that must be taken into account to understand its role in our lives.

The intentional connection is necessary because it is the only way in which the two contradictory beliefs are linked *as contradictory*, since the 'direct' rational links between incipient and final belief state are severed and their contents are *prima facie* independent of each other. Once the intention is introduced, the contradictory beliefs, though not linked in accordance with ordinary rationality requirements, would nonetheless be related on account of their content via the intermediary of the intentional action. Intentionalism demands that there is a practical syllogism leading from the belief p to the belief not-p. To the extent that intentions are linked to reasons, the practical syllogism must take into account the content of p and not-p. Yet insofar as the beliefs must be kept apart to avoid the agent's realising the contradiction, the intention must be cast in such terms as to avoid bringing the beliefs together or bridging the mental boundary.

For Davidson, an action is intentional only if it can be rationalised by the agent, i.e. if there is a combination of beliefs and pro-attitudes (such as desires, evaluations, etc.) in the light of which performing the action is reasonable: "...if

[57] Ibid., p.145

[58] Indeed, the term 'self-deception' is first recorded in early Christian writings and came again to the fore in the thought of English protestants like Butler; cf. Steffen (1986), Martin (1986), and Holton (2000).

[59] We will see later that although this may be Davidson's reason, it cannot be the right explanation for the moral dimension of self-deception.

someone acts with an intention, he must have attitudes and beliefs from which, had he been aware of them and had the time, he could have reasoned that his action was desirable... ."[60]

Davidson's characterisation of the intentional action in terms of 'producing a belief in the negation of p' leaves room for competing interpretations: either the agent intends to produce the belief by ordinary, rational means, or her action is performed with the intention of deceiving herself. But Davidson explicates elsewhere that "The self-deceiver must intend the 'deception'"[61], because "it is not self-deception simply to do something intentionally with the consequence that one is deceived, for then a person would be self-deceived if he read and believed a false report in a newspaper."[62] So we can safely conclude that Davidson thinks the agent who deceives herself acts in the knowledge that the belief she hopes to eventually acquire will not be the result of ordinary, i.e. rational, processes of belief-formation.

This also explains Davidson's position relative to that of other theorists of self-deception: many of those who label themselves 'anti-intentionalists' and oppose Davidson would not deny that intentional action is involved somewhere in the aetiology of self-deception. For example, they would permit that the agent intentionally looks for new evidence, and that her 'looking for evidence' is biased; but they would object to the suggestion that the agent acts in a way that is 'intentionally biased'.

If self-deception is 'intentional', we attribute to the agent the formation of a practical syllogism involving a pro-attitude and a belief in the light of which the action is reasonable. Davidson mentions in passing what kind of pro-attitude could play the part: "...while the self-deceiver may be motivated by a desire to believe what he wishes were the case there are many other possibilities. Indeed, it is hard to say what the relation must be between the motive someone has who deceives himself and the specific alteration in the belief he works in himself."[63] This anodyne description of the practical syllogism is not just an editorial shortcoming, but an indication of a serious difficulty Davidson fails to appreciate – that the obstacle he perceived at the start of his project, viz. that explanations of self-deception tend to strip it of its irrational character, has re-emerged. If self-deception is intentional, its explanation must invoke beliefs, desires, etc.; yet this is precisely what Davidson

[60] IN, p.85
[61] DD, p.144
[62] Ibid., p.144

wishes to avoid to make plausible his contention that self-deception is inherently irrational.

5. The Pro-Attitude and Lazar's Criticism

What kind of belief-desire-complex could rationalise the action performed under the description 'getting myself to believe p in the face of overwhelming counterevidence'? It would have to involve some pro-attitude along the lines of 'I want to believe that p', or 'Life would be so much nicer if I believed that p',[64] and a belief like 'I could believe that p if I ignored the counterevidence, or if I paid more than proportionate attention to the evidence in favour of p'. Together with the belief that by paying attention to some parts of the evidence and not others I can deceive myself and acquire the belief p, the pro-attitude yields a practical syllogism that makes my self-deception intentional. In a simple case, the agent holds evidence for p; she has a pro-attitude to holding the belief that not-p.[65] So she acts so as to acquire the belief not-p: she looks for evidence that would ordinarily count as evidence for not-p, she avoids looking at what could constitute evidence for p.

But, as Ariela Lazar notes, "the intention to form a belief in itself, forms an obstacle towards the formation of the irrational belief."[66] That is, whatever the final tally of the evidence so acquired, the agent who has intentionally acquired this evidence in order to get herself to believe a falsehood knows that this evidence is not trustworthy. It is not 'good' evidence that can be relied on to form a true belief; so she cannot genuinely form the belief not-p. The intentional action that, according to Davidson, occurs in self-deception must involve something like a shifting of attention, or a manipulation of evidence in a way that the self-deceiver expects will affect her conclusion. Yet if she intentionally shifts attention away from some evidence, or attends to the evidence that supports her pet hypothesis very early or very late in her investigation because she knows that people tend to give more weight to this evidence than to evidence that they encounter towards the middle of their investigation, then these strategies, *because* the agent formed an intention to use them, are unlikely to be successful.

---

[63] Ibid., p.144

[64] We should note that the pro-attitudes cited all involve, in one form or another, a conception of oneself as 'the kind of person who believes not-p'. The agent conceives of herself as a psychological object and assesses her own state of mind, especially the set of beliefs she holds, from the outside.

[65] She does not have a pro-attitude to its being the case that not-p; or at least this pro-attitude cannot explain the self-deception, but only her acting to bring about the state of affairs not-p.

[66] Lazar (1998), p.25

Beliefs, by their very nature, are world-guided; we assume our beliefs are true to the world, and if we doubt that they are, we discard them. Davidson shares this view of beliefs, and this is reflected in his account of interpreting others against the background of what we take to be true.[67] Beliefs are propositional attitudes, and their content is normatively linked to the content of other propositional attitudes. This does not mean that they are not also causally linked to the world or other beliefs; but *as beliefs* they have content that has rational force and is subject to normative constraints. By contrast, the action involved in Davidson's account of self-deception plays on the purely causal connections of beliefs to other beliefs and the world, and has the explicit aim of overpowering the rational forces at work in the formation of beliefs. Knowing that one's belief was formed not in accordance with the actual state of the world, but is instead the result of purely causal (non-rational) forces, undermines one's trust in the belief *as a belief*, i.e. as a propositional attitude with a content that normatively depends on how the world is. The belief must therefore be discounted.

The requirement imposed upon one's own beliefs – that they are well-founded and do not contradict the actual state of the world – is not, as one may be tempted to think, deactivated when one deceives oneself. This becomes obvious when we consider an important feature of self-deception that Davidson does not explicitly address: 'rationalisation', or – to distinguish it from Davidson's use of the same term in the context of action explanation – 'false rationalisation'.[68] Someone who deceives herself does not usually admit, either to herself or others, that her beliefs or actions are based on no or only partial evidence; someone who deceives herself would not ordinarily admit that she were holding contradictory beliefs if someone pointed it out to her.[69] This is indeed one of the most obvious distinctions between those contradictory beliefs involved in self-deception, and those that are merely due to a lack of reflection and are set straight when pointed out to the agent (what we called 'innocuous contradictory beliefs'). The agent who deceives herself tries to defend her position with reasons for her holding the belief, and the only reason for believing p is evidence for p. It is a common observation that someone who we deem to be deceiving herself will insist that she is completely rational, and provide more or less reasonable, though ultimately unconvincing, explanations of why she

[67] Cf. sections 1 and 2 in Ch.1
[68] Cf. Sanford (1988)

is justified in holding the belief in question – unconvincing, because they cannot destroy our impression that the *real* reason for her holding this belief is a different one. If she did not see herself as someone who holds beliefs that are rational, and holds them because she has good evidence for them, false rationalisation could not be explained.

Thus, from Lazar's criticism, we must conclude that an intention to form a false belief is self-defeating, because beliefs just are the kinds of things about which we think they come into being independently of our preferences for or against them.[70]

6. Possible Ways Around Lazar

This is a powerful argument, and we must take care to avoid the difficulties it indicates. However, its success rests on the claim that we know our intentions. If the agent is not aware of the intentions in the first place, she may not have the opportunity to link her knowledge that the evidence is skewed with the belief the evidence seems to warrant, and the belief would not need to be disavowed.

The difficulty arises because we must satisfy three demands at once: we must locate the intentional connection in our ordinary framework of mental explanation; we must keep apart the two contradictory beliefs by means of the mental boundary; and finally we must hold apart the belief not-p and the intention to form it if the belief is to play its ordinary role (which must be required as part of our conception of self-deception as involving ordinary beliefs). If the agent were able to cross reference the thought that her forming the false belief depends on the practical reasoning that guides intentional self-deception (rather than the rational warrant provided by the evidence acquired) with the false belief not-p, she would have to disavow this belief. If agents recognise that the link between belief and truth that the concept of belief requires has been cut off, they cannot keep thinking of their state of mind as a belief, but at best as a wish.

Davidson's first option is to offload the difficulties into the realm of the unconscious and argue that the agent need not be aware of the intention at all. But two considerations make this strategy unattractive: first, introducing unconscious mental states into the account has little explanatory value and is, at this stage in

---

[69] It seems as if Davidson thinks that she would admit it; or that is at least how I read his claim that the agent who deceives herself is in an 'unstable' state. Cf. DD, p.140

[70] This is, obviously, intended to be on the same level as the statement that if we deceive ourselves, we form beliefs because we want to form them; it does not even touch on questions about, say, pragmatism, which, although sometimes verbally close, are not our concern here.

Davidson's argument at least, pure stipulation. Second, unconscious beliefs, desires, and other mental states ought not to be conceived of as unrelated to the rest of the mind. They have their usual explanatory power because they have the role of ordinary mental states in many contexts, but are not easily accessible to the agent. There is thus no reason to believe that unconscious beliefs could not impede the formation of beliefs they obviously contradict.

A more attractive way out of this conundrum is for Davidson to use his own 'conceptual tool' – the mental boundary – to overcome the difficulties of intentionalism. He explicitly denies that one part of the partitioned mind can be identified with the 'unconscious', and can thereby avoid the criticism advanced in the previous paragraph. The power of this move thus depends on how convincing the stipulation of the mental boundary is, and whether the intention can be located 'behind the boundary' without substantial difficulties or explanatory awkwardness. If the intention to influence a belief can somehow be shielded by the mental boundary from the belief that is produced (and assumed to involve all the commitment beliefs ordinarily involve) then Davidson's account of self-deception would succeed. To assess this possibility, we must consider the role of the mental boundary as a way of talking about contradictory beliefs that the agent holds yet does not 'put together' (in the sense in which we say 'She does not put two and two together'). The agent does not put these beliefs together because or insofar as – it is not clear how much of this is causal, and how much is a matter of logic – they are on different sides of the mental boundary; in the case we are concerned with, she does not put together her knowledge of the intention to manipulate the available evidence and her belief based on this (manipulated) evidence.

So the most straightforward solution seems to locate the intention behind the mental boundary, together with the belief p that causes the belief not-p. But this would isolate the intention from the action to which it belongs, leading to the very same problem we encountered earlier: the relations between the events located on different sides of the mental boundary are for Davidson purely causal, yet the link between a practical syllogism and the action that it rationalises must be 'rational'. Locating the practical syllogism and the action on one, the belief that is caused by the action on the other side of the mental boundary entails that we can identify the links between them only as a causal connection, and not as that between an intention and the action it is the intention to perform. With the rational connections

severed, the practical syllogism cannot rationalise the belief formation, and we cannot talk about 'intentional action'. This is, in fact, an argument against any strategy that tries to shelter the intention from the acquisition of evidence the intention brings about: to the extent that relations between an intention and the intended action are subject to requirements of rationality and reason, the very link that makes the intentional action possible entails that an intention to form false beliefs is *prima facie* self-defeating.

7. Mental and Physical

Davidson's account of self-deception as dependent on the mental boundary is unsatisfactory; in fact, the very notion of 'mental boundary' seems suspect in the light of the considerations developed just now. But before we reject Davidson's account completely, we should ask how, if at all, it is possible to hold overwhelming evidence for p, yet acquire the belief not-p, i.e. how one can ignore parts or all of one's evidence for a conclusion. Davidson's answer is: by ignoring the requirement of total evidence.[71] The earlier quote from Davidson illustrates his explanation for the failure to apply the principle in question:

> What causes [the requirement of total evidence] to be thus temporarily exiled or isolated is, of course, the desire to avoid accepting what the requirement counsels. But this cannot be a reason for neglecting the requirement. Nothing can be viewed as a good reason for failing to reason according to one's best standards of rationality.
> In the extreme case, when the motive for self-deception springs from a belief that directly contradicts the belief that is induced, the original and motivating belief must be placed out of bounds along with the requirement of total evidence. But being out of bounds does not make the exiled thought powerless; on the contrary, since reason has no jurisdiction across the boundary.[72]

So the desire to avoid accepting what the requirement counsels causes the failure to accept what it counsels. Yet for obvious reasons this description of the problem does not fit into the picture of mind and psychology that we found in Davidson's account: although there is a causal, there is no rational link between the desire and what it causes, because plainly for Davidson there can be no reason to avoid what the requirement counsels. Thus, ordinary reason explanation – involving pro-attitudes and beliefs – cannot be used here. From Davidson's discussion it seems to follow that any explanation of how the alleged desire to ignore the principle affects

---

[71] There is a certain temptation to infer from Davidson's claim that weakness of the warrant is somewhat similar to weakness of the will that Davidson's account in WW might help us here. But this ignores an important different between the two 'weaknesses': there is no obvious difficulty in thinking that an agent can act while believing that her action is a result of weakness of the will - though there are less obvious ones; cf. Belgum (1978), ch.3 - but there are obvious difficulties in thinking that an agent can hold a belief and think that it is not warranted by how the world is.
[72] DD, p.148

the process of belief formation must makes use of purely physical (non-rational) terms.

But if no explanation in terms of reasons is available, self-deception must ultimately be explained in terms of physical cause and effect, which is not what our psychological vernacular suggests.[73] Davidson's argument for the anomalousness of the mental, viz. that there is no type-identity between mental and physical events, prevents any explanation from the physical base to the supervenient mental structure, and thus also an explanation of the causal link between the desire and the effects it has on the formation of beliefs. A mental explanation is successful, Davidson says, because we assume it is supported by a physical explanation; but this cannot solve the problem at hand, since this only applies to ordinary mental explanation, which is taken to be successful because there are underlying causal forces that could be explained in terms of physical laws and events. We cannot, by contrast, find an explanation in terms of reasons on the basis of a given physical, non-reason explanation. At the very foundations of Davidson's account of self-deception lies therefore the implicit claim that self-deception can only be made sense of at the physical level. As this does not fit our ordinary scheme of mental explanation, it cannot explain the vernacular – psychological – notion of self-deception we are after.

8. Conclusion

Davidson's account cannot make clear how talk of self-deception fits into his picture of the mental as governed by rationality, yet dependent on the causal connections between events. Despite the introduction of the mental boundary that permits us to disentangle causal and rational connections in the search for an explanation of self-deception, Davidson's account cannot accommodate our ordinary talk about the phenomenon, which makes sense of it while simultaneously recognising its irrational character. The introduction of intentional action, though initially appearing as a useful device for overcoming the problems of Davidson's account, eventually proves prone to the same difficulties: it cannot accommodate the sub-rational, yet personal-level mechanism that must be at work if our ordinary conception of self-deception is appropriate. We are thus left with the choice

[73] This criticism of Davidson is not to be confused with the charge of epiphenomenalism as discussed, e.g., in McLaughlin (1993); in particular, it does not suffice to give up the nomological view of causal explanation to solve this problem.

between two unacceptable options: either self-deception is explained in terms of intentions, which allows us to account for the place it has in our psychological vernacular but violates the commitment to the notion of rationality that characterises Davidson's picture of the mind; or the explanation of self-deception requires reference to events described in physical rather than mental terms, leaving us with a picture of self-deception according to which it is nothing but a mental short-circuit. In the following chapter, we shall see that this conclusion is not just the result of a narrow reading of Davidson, but holds also true when we approach the problem from a different angle and discover that there is no room for self-deception in Davidson's account of mental concepts.

Third Chapter: Mental Concepts

In the previous chapter I discussed Davidson's account of self-deception, and showed that it falls short of a satisfying explanation of the phenomenon. In this chapter, I raise a question that gets closer to the foundations of Davidson's philosophy: where can the concept of self-deception fit into Davidson's general account of the mind?

The realm of the mental is, for Davidson, delineated by the use of a particular vocabulary, defined by the 'essential use of intensional terms' and employed in interpreting others. To elucidate the parameters of the mental, we must analyse this vocabulary. In the previous chapter, we saw that Davidson's account cannot make sense of self-deception without undermining its irrational character, because the explanatory framework admitted by Davidson cannot accommodate the sub-rational, yet personal-level mechanism that must be at work if our ordinary conception of self-deception is appropriate. In this chapter, a similar conclusion will be reached; but this time it follows from the interpretationist foundations of Davidson's philosophy of mind. The groundwork for this result is laid in the first part, which argues that interpretation theory should explain the univocality of mental concepts in terms of 'imaginative projection' if it is to accommodate Davidson's claim that the rationality against which we measure others is always our own. The second part of the chapter shows that such a view makes it impossible to account for the concept of self-deception: if interpretation, i.e. projective imagination, is constitutive for mental concepts, and we cannot come up with self-deception in projective imagination, the concept of self-deception begins to look problematic.

1. Interpretation Theory and Univocality

Davidson thinks of his interpretation theory as a rational reconstruction of the ordinary practice of attributing mental states, as well as a constitutive account of the meaning of mental terms.[74] It must therefore also be subject to the requirement of 'univocality'[75]: mental terms, or concepts, have a third- as well as a first-personal use, and they must have the same meaning in both cases. If mental terms were not univocal, the content of my claim 'I believe the sun is shining' would not be related

[74] Cf. sections 1 and 2 in Ch.1.

in the appropriate fashion to another person's statement about me (e.g. 'He believes the sun is shining') to enable us to talk about each other's mental states. Denying univocality leads to either solipsism – claiming that other people experience nothing that is like 'my pain', so that my pain is the only pain around – or the assertion that the first-personal use of mental terms is not properly psychological, but only 'presentational', i.e. it presents an assertion without contributing to its content.

But the requirement of univocality ought not to relieve theories of the mental of accounting for the *prima facie* differences between first- and third-person stances: an agent usually becomes aware of her beliefs and mental states in ways very different from how other people become aware of them, and her claims to knowing her own states do not require the same kind of justification knowledge claims about other people's mental states demand. Such differences have been denied, most prominently by logical behaviourists[76]; but acknowledging their existence seems more sensible: although the denial stems partly from a meritorious rebuttal of the idea that one has infallible knowledge of one's mind – which would leave no room at all for self-deception and other forms of irrationality – it ignores the possibility of accommodating fallibility while permitting for differences between first- and third-personal stances in psychology, and it puts a serious strain on an account of interpersonal interpretation as envisaged by Davidson.

Interpretation theory appears to inherently possess a third-personal slant: it assumes that interpretation is constitutive for mental concepts, and at least on the surface it is plausible to suppose that the outcome of any act of interpretation is first and foremost determined by the interpreter's imposition of meaning upon the behaviour of the interpreted subject. There are two ways of accommodating the first-personal use of mental terms in this picture. Either we could assume that it requires different mental concepts from those used in interpreting others; or we could assimilate first- and third-personal uses and insist that they are simply two different *uses* of the same concept. Since the former amounts to a denial of univocality, we ought to pursue the second option. But this is not without difficulties of its own: it may appear to commit us to redescribing the first personal use of mental concepts in terms familiar to us from interpersonal interpretation. Thus our first task is to show that it is possible to side-step the problematic interpretation of Davidson's philosophy that makes the first-personal use of mental

[75] See Strawson (1959), part I, ch.3, for the source of this requirement in contemporary discussions.

terms secondary to their third-personal use, and to provide an account of interpretation theory that respects univocality as well as the other commitments Davidson's philosophy undertakes.

2. Instrumentalist Interpretations

Davidson does not spell out in much detail why we use mental terms at all. Yet this is a pressing question, because other philosophers who have adopted Davidson's interpretation theory, most prominently Daniel Dennett, explicitly recommend an instrumentalist conception of intentional psychology. For Dennett, intentional psychology deals with the states of an intentional system, defined as "a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches, ...)"[77]. It is adopted because it is useful in predicting and explaining the behaviour of a system that would otherwise be too complex for us to cope with. If we are to assume that Davidson follows this lead and takes the identity of mental states to be constituted by their role in the explanation and prediction of behaviour, then the univocality of first- and third-personal mental terms would require that in 'self-interpretation' – the use of mental terms in the first person – the agent ascribes mental states to herself according to their role in predicting and explaining behaviour. Conceiving of the first-personal use of mental terms in close analogy to the third-personal use not only ascertains their univocality, but also fits the observation that first- and third-personal ascriptions of mental states usually agree in their results. This fact would now have a simple explanation: first- and third-personal ascriptions agree because they use the same concepts and proceed by the same methods, those known from interpersonal interpretation.

But reading Davidson's interpretation theory in this instrumentalist vein creates serious difficulties. It fails to make sense of divergences between first- and third-personal uses of mental terms, the most obvious – and the most directly relevant to interpretation – concerns the different attitudes we can ('grammatically', in Wittgenstein's sense) take towards our own beliefs and those of others: I can take other people's current beliefs (and their past beliefs, and my own past beliefs) but

[76] Most famously perhaps by Gilbert Ryle; see, e.g., Ryle (1994), especially p.20.
[77] Dennett (1971), p.87; for a further development of this view, see Dennett (1991).

not my own current beliefs to be false.[78] If we take this into account, we must reconsider the function of interpretation in first- and third-personal cases. When the interpreter who attributes mental states to someone else finds that her attribution comes out as inconsistent, or involves attributing beliefs that are out of line with how the world is, she has three options[79]: first, she can revise the attribution of beliefs, and hope to achieve a higher degree of consistency, or a better fit with the world. Second, if full consistency between beliefs and the world is not achievable – and this will frequently be the case – she can settle with ascribing beliefs that are false, and accept that the subject got some things wrong. Third, the interpreter has the alternative option – perhaps less frequently noticed in philosophy than chosen in life – of changing her own view of the world and accepting that the subject got it right after all.

If the agent were to 'interpret herself' – that is, apply the strategy used in identifying the mental states of others to herself – these three options shrink to one. The first option, 'revising the attribution of beliefs', takes on a different meaning when applied to oneself: if the interpretation of one's belief is altered, the identity of the belief that is being interpreted must change too: beliefs have, according to Davidson, no identity independently of interpretation. If we could meaningfully say that the belief had stayed the same, and only the interpretation of it had changed, we would fail to take seriously the idea that the interpretation involved here is constitutive. 'Self-interpretation', in this sense, must be distinguished from cases in which interpretation is what we might call 'investigative', which is the ordinary notion of interpretation outside of the philosophical context of interpretation theory: I can, for example, think very hard about my relationship to my siblings and come to the conclusion that, properly interpreted, my behaviour in their presence reveals serious resentment. This kind of self-interpretation relies on the idea that the beliefs or emotions are not constituted but discovered by self-interpretation. If it were otherwise, 'self-interpretation' in this second sense would be thoroughly unattractive: why would I wish to *bring about* a feeling of resentment in me that did not exist before? We must thus distinguish the 'investigative' sense of 'self-interpretation' from the sense relevant for our discussion; and in the case of 'constitutive' self-interpretation, changing one's interpretation of one's own belief

[78] Cf. Wittgenstein (1953), p.190: "If there were a verb meaning 'to believe falsely', it would not have any significant first-person present indicative."
[79] Cf. Moran (1994), pp.165ff.

amounts to changing the belief. Thus, the first option available in third-personal interpretation collapses into the third.

The second move, ascribing false beliefs to oneself, is not an option for the self-interpreter either.[80] Explaining why this is true is much harder than recognising that it is, and for the purpose of this discussion I will settle with the latter, and merely describe the phenomenon in some more detail. The differences between first- and third-personal stances are illustrated by the alternative attitudes taken towards the propositions 'p' and 'I believe p'. From the third-personal stance, there is a clear distinction between ascribing to the agent the belief 'p' and holding p to be true, while from the first-personal perspective, they appear, although theoretically distinct, as practically equivalent. The case is even clearer for 'I believe p' and 'not-p'. From the first-personal perspective, these propositions "are not in logical contradiction with each other, yet they do systematically conflict with each other, though their counterparts in the mind of the interpreter do not. ... [The self-interpreter] does not have the option of treating the relation between his first- and second-order beliefs as a conflict between two different belief systems, but only as a conflict within *one* view of the world."[81] This suggests that self-interpretation and the use of mental terms in the first person cannot merely be an explanatory or predictive device; the agent cannot just want to predict what she believes, but actually aims at believing the truth.[82]

Thus, once we take into account the differences between the stance we can take towards our own mental states and the stance we can take towards those of others, and incorporate these differences into our understanding of interpretation, we recognise that interpretation theory cannot depend solely on prediction and explanation: mental concepts, when applied in the first person, depend on the agent's capacity for revising her beliefs, making up her mind, or, more generally, deliberation.

---

[80] This obviously only captures the standard case of the ascription of currently held beliefs to oneself identified directly by use of the term 'I', and not cases of ascribing past beliefs to oneself or ascribing beliefs without recognising that the subject holding these beliefs is oneself; cf. Perry (1979).

[81] Moran (1994), p.167; emphasis in the original.

[82] This mirrors the second tenet of Davidson's holism, which specifies that the notions of belief and truth are intrinsically connected. Cf. Ch.1, section 1

3. The Deliberative Alternative

By examining the fundamental distinction between self-interpretation and the interpretation of others, we elucidate the differences between first- and third-personal uses of mental terms: from the first-personal perspective the question about what the agent believes amounts to a question about what is the case, while from the third-personal perspective there is no such 'transparency'[83]. This observation is captured in the distinction between a theoretical and a deliberative stance towards beliefs. The deliberative question asks 'What ought I to believe?', while the theoretical question, when it is at all asked about oneself, takes the form 'What am I likely to believe?'[84] The distinction between deliberative and theoretical question should not overshadow their interdependence; for example, in the first-person case, deliberative and theoretical questions are fundamentally entangled – the theoretical question 'What am I likely to do?', or 'What am I likely to believe?', cannot ignore the corresponding deliberative question.

The emphasis on deliberation fits well into Davidson's account of the asymmetries between first and third person, as laid out in his 'First Person Authority'. There, Davidson's focus is on how speaker and interpreter respectively determine the meaning of the speaker's utterances. The basic set-up is this: when someone says 'Wagner died happy', interpreter and speaker alike can know that the speaker holds this sentence true on this occasion, in which case an interpreter who knows the meaning of the words uttered also knows what the speaker believes: "...you and I both know that I held the sentence 'Wagner died happy' to be a true sentence when I uttered it; and that I knew what that sentence meant on the occasion of its utterance. And now there is this difference between us, which is what was to be explained: on these assumptions, I know what I believe, while you may not."[85] However diligently the interpreter works on an adequate interpretation of the speaker's words, taking into account all available clues, she is subject to error. The speaker can also be mistaken about the meaning of his own words, because "what his words mean depends in part on the clues to interpretation he has

---

[83] Cf. Evans (1982), pp.225ff.

[84] The theoretical question can be asked about oneself; but it is not asked very frequently. A famous example is offered by Sartre: the gambler decides not to gamble again, but then begins to wonder how likely it is that he will stick to this decision given his history of backsliding and going back to the gambling table. See Sartre (1956), pp.70ff., quoted in Moran (2001), p.79. The distinction between 'theoretical' and 'deliberative' question also reflects that between 'transcendent' and 'thinned-down' views of rationality in Ch.1, section 5.

[85] FPA, p.12; also cf. KOM, pp.36ff.

given the interpreter, or other evidence that he justifiably believes the interpreter has."[86] Nonetheless there is a fundamental asymmetry "that rests on the fact that the interpreter must, while the speaker doesn't, rely on what, if it were made explicit, would be a difficult inference in interpreting the speaker."[87]

For Davidson, the underlying asymmetry is essential to interpretation, and he analyses it in terms of the difference between the position one is in when interpreting one's own words and the position one is in when analysing someone else's words: "The speaker, after bending whatever knowledge and craft he can to the task of saying what his words mean, cannot improve on the following sort of statement: 'My utterance of "Wagner died happy" is true if and only if Wagner died happy.' An interpreter has no reason to assume that that will be *his* best way of stating the truth conditions of the speaker's utterance."[88] From the first-personal perspective, our actions – linguistic and non-linguistic – are chosen on the basis of deliberation, leading to what we think is the right thing to do or say; but from the third-personal perspective, the speaker's deliberation is not the end, but the beginning of the matter, since understanding requires that the interpreter identifies the beliefs and other mental states that lead to the particular actions of the speaker.[89] These arguments by no means prove that Davidson must accept an account of mental concepts that emphasises deliberation instead of prediction and explanation; but it shows that such a reading of Davidson's interpretation theory fits into his wider philosophical commitments and can avoid problems the alternative reading creates.

4. Deliberation and Third-Personal Concepts

The introduction of the deliberative stance makes room for an alternative account, which avoids instrumentalism about mental terms and need not give primacy to either the first- or the third-personal use of mental concepts. It also proves its worth as an interpretation of Davidson's philosophy by accommodating certain strands of his account of radical interpretation that would otherwise appear odd, especially the claim that the interpreter's task of attributing maximally rational beliefs is

---

[86] FPA, p.13
[87] Ibid., p.13
[88] Ibid., p.13; emphasis in the original
[89] Critics of Davidson's position tend to overlook this distinction, and instead insist that Davidson's holism and interpretation theory require that the speaker must somehow aim at consistency, rather than at getting things right. See, e.g., Hamilton (2000), especially p.28.

equivalent to maximising the agreement between herself and the subject of interpretation. Interpretation requires – or is equivalent to – imposing one's own standards of rationality onto what other people do, because it demands redescribing what they do in ways that make their linguistic and non-linguistic behaviour come out as rational, and the only standard of rationality we have available, Davidson asserts over and over again, is our own.

On the suggested reading, Davidson's contention that 'maximising agreement' and 'maximising rationality' are equipollent does not leaves us with a free choice between using one or the other strategy; instead, the two simply amount to the same thing: to maximise an agent's rationality in the process of attributing beliefs, desires, and other intentional mental states requires describing what she thinks in terms that are as close as possible to what the interpreter thinks. The standard of rationality held by the interpreter is reflected in her own beliefs – that is simply another way of putting the observation that one cannot ascribe to oneself false beliefs; and the standard of rationality she holds cannot be spelled out or codified independently of these beliefs, for the reasons discussed in chapter 1. If, as instrumentalism requires, the agent were to apply either a theory or some general principles of rationality to herself, she would have to appeal to some pre-existing notion of rationality to apply the theory or principles correctly. Similarly, interpretation of others cannot be fixed by rules which are independent of the interpreter's rationality, because the application – the meaning – of these rules would depend on the prior understanding of rationality the interpreter possesses.

If we emphasise deliberation and projective imagination in our account of interpretation, we ought to picture third-personal interpretation as transforming the theoretical question 'What does she believe about p?' into the first-personal question 'What would I believe about p, were I in her position?' that entails the explicitly deliberative question 'If I were in her position, what should I believe about p?'. Interpretation is then a matter of imagining being in someone else's intentional mental states, given the interpreter's knowledge about that person's current states, actions, and position in the world. Univocality of mental concepts flows from our capacity to reason counterfactually and from our shared commitment to rationality – the ability to reason, to justify, to act in accordance

with what one takes to be justified etc. – which informs our understanding of each other and the deliberative stance we take in our own lives.[90]

What differs between deliberation and projective imagination are the inputs to the system, e.g. the location of the agent in space and time, and the impact this has on what she perceives, believes, etc.; in particular, some of the beliefs imputed to others can be false. While beliefs can explain actions even if they are false, they can justify actions only if true.[91] The justificatory project inherent in the first-personal perspective requires that false beliefs be eradicated. Interpreting others, by contrast, does not require such a radical cure: we can assume that others hold false beliefs without thereby undermining our capacity to imagine their justification, so we can explain someone else's behaviour while being aware of the falsehood of some of her beliefs. This does not undermine the univocality of mental terms, not even that of 'belief', despite the apparent discrepancy between its first- and third-personal uses: the first-personal use of 'belief' is not solipsistic, i.e. 'I believe p' leaves room for the distinction between beliefs and the world. The differences in use do not require distinguishing between the concept used in the first and the third person, but can be attributed to the different interests with which it is employed in each case – deliberating vs. imagining deliberating to understand others.

In addition to the obvious merits an account that emphasises deliberation and projective imagination has for understanding Davidson's characterisation of interpretation in terms of 'maximising agreement', it alleviates a worry that all interpretation theories face: that our interpretation is mere imposition, a form of mental colonialism. On the proposed understanding of Davidson's theory, when other people talk about their own beliefs and thoughts and other attitudes, they make use of the deliberative stance and display their commitment to rationality. So the interpreter's use of rationality becomes less a matter of imposing something that was not there and more a matter of retracing what is already present. Finally, spelling out Davidson's theory in terms of projective imagination enables us to avoid the hint of instrumentalism that Dennett brought onto the stage: his view could only accommodate the demands of univocality by unduly emphasising the third- over the first-personal use of mental concepts, and by ignoring the difficulties that arise when a 'theory of rationality' is summoned. Since the overall thrust of

[90] See Heal (1995), esp. pp.52ff., for a related 'replicationist' argument; the link between holism and replicationism or projective imagination is made very clear here.

Davidson's philosophy points away from instrumentalism, a reading of interpretation theory that invokes deliberation and projective imagination is at the very least congenial to Davidson's view.

5. Interpretation and Self-Deception

Our conclusion so far is that the best available reading of Davidson's philosophy requires linking interpretation to projective imagination. But in spite of its advantages, such a reading also brings to light important problems in Davidson's theory, which are linked to the difficulty of finding a proper place for the concept of self-deception. If projective imagination is at the basis of our understanding others, then we would expect that the attribution of self-deception is also the result of deliberating 'through someone else's eyes'. But it seems impossible to make a coherent or stable attribution of self-deception to oneself; so it becomes difficult to see how we can imagine this state in others. As we cannot know what it would be like for ourselves, we cannot – or so it seems – know what it would be like for others to deceive themselves. If understanding others requires imagining having their mental states on the basis of our knowledge of their actions and the shared ideal of rationality, and we cannot make sense of the idea that we are deceiving ourselves about a certain topic, then we cannot attribute self-deception to others in the same way in which we attribute innocuous beliefs or desires.

The first step towards solving the problem requires identifying the precise limits of our capacity to ascribe self-deception to ourselves. This is the aim of the following section. In the subsequent section, I analyse what might at first seem like a parallel case, the attribution of false beliefs to someone, and argue that there are important differences between these cases. These differences, and how they make attributing self-deception by projective imagination – and thus finding a proper place for the concept of self-deception in Davidson's picture of the mind – impossible, are the topic of the final section.

6. Self-Attributing Self-Deception

Stevens, the butler in 'The Remains of the Day'[92], slowly and painfully works trough his memories of the inter-war period when he was employed by the

[91] Cf. ARC, p.8: "Your stepping on my toes neither explains nor justifies my stepping on yours toes unless I believe you stepped on my toes, but the belief alone, true or false, explains my action."
[92] Ishiguro (1989)

aristocratic amateur diplomat Lord Darlington, one of the chief proponents of 'Appeasement'. Stevens sees these years (at least sometimes) as the pinnacle of his success as a butler: he was the man who ventured from the shadows to ensure that the events that (Stevens thought) determined the fate of the world ran smoothly, and that powerful men got their cup of tea on time. After the war, Stevens on several occasions denies knowing Darlington – dead, and indicted by public opinion for his prominent role in the ill-fated project of 'Appeasement' – yet does not acknowledge that he is ashamed of his former employer.

I am quite certain that Stevens deceives himself here about his attitude to Lord Darlington, and this judgement is perhaps shared by most readers. But it would be inconceivable for Stevens to come to this conclusion. If Stevens thought something along these lines: 'I have come to believe that I deceive myself when I think that I am not ashamed of my work for Lord Darlington', there would be no obvious way to understand this. Interpreting it requires at the very least that we take Stevens to misuse the present tense in describing what he deceives himself about. There are more or less plausible ways to understand his sentence; but we cannot take it at face value and interpret his words to have the same meaning as they would if someone else said about Stevens that he is deceiving himself.

7. False Beliefs and Self-Deception

It might seem obvious that the reason for this difficulty is simply this: when we charge Stevens with self-deception, we imply that he holds a false belief.[93] If Stevens were to express the belief that he deceives himself, he would say that he holds a false belief about the subject-matter of his self-deception. Our ordinary way of understanding such a claim is that he must have relinquished his false belief. This is simply part of the deliberative stance we discussed earlier: when the agent discovers that one of her beliefs is false, she must change her view of the matter, since she can neither settle with attributing to herself a (current) false belief, nor simply reinterpret the meaning of her belief without thereby also changing its identity.[94]

---

[93] Some theorists of self-deception think that even this is an excessive commitment; according to them, self-deception does not require a false belief but merely an internal inconsistency; see, e.g., Scott-Kakuris (1996). I believe that when this inconsistency is spelled out properly, it requires attributing some false belief; but proving this is beyond the scope of this thesis, and irrelevant for the development of the argument.

[94] This does not exclude the possibility of thinking that one is deceiving oneself now with respect to *some* belief; when I have spent enough time writing about self-deception, I am pretty certain that

But if this were all there is to the difficulties of self-attributing self-deception, and attributing self-deception ran parallel all the way to attributing false beliefs to oneself, our argument would face a problem: it would seem that the difficulties raised by self-deception cannot be real, or that, if they are, Davidson's account is flawed beyond repair. The first horn of the dilemma is this: if attributing self-deception to oneself poses the alleged difficulties, and is thus excluded from the realm of mental concepts circumscribed by Davidson's interpretation theory, then, assuming that self-deception and false belief are similar concepts, we ought to be equally incapable of making sense of the concept of false beliefs. This amounts to a *reductio ad absurdum*, either of Davidson's entire account or of my reading of it in terms of projective imagination. The second horn of the dilemma is equally forbidding: if there is no *reductio*, because the notion of false belief fits into Davidson's account of mental concepts, then there ought to be no difficulty with the concept of self-deception either, since they are similar. So our criticism of Davidson would be mistaken.

The way out of this conundrum is to deny what might seem like common sense: that the concept of self-deception is analogous to that of false belief. It becomes clear that drawing this parallel is too hasty once we attend to the different roles the concepts of false belief and of self-deception play in our mental economy.

The notion of false belief is logical, since a false belief has the same content – and plays the same psychological role – as a true belief. As it is not psychologically differentiated from a true belief, how someone comes to hold a false belief does not *per se* require explanation, at least no more so than how she comes to hold a true belief. The notion of self-deception, by contrast, not only entails the falsehood of the belief at hand; it also invokes an explanation of the ways the false belief is acquired. Charging someone with holding a false belief is simply pointing out a discrepancy between what she thinks and what is the case; charging someone with self-deception, by contrast, amounts to saying that she holds a false belief, *and* that she holds that belief for a reason, or as a result of a particular action.[95]

---

everyone is deceived about some S. But I would not be able to fill in a particular subject-matter for S, since if I genuinely thought I were deceiving myself about that S, I would check my beliefs about S, either to put them in order if they are messed up, or to conclude that nothing is wrong with them.

[95] For a detailed account of the intricacies of self-deception, see the brilliant van Fraasen (1988). It shows that less sophisticated accounts, like Beyer's (1998), which do not distinguish between the difficulties posed by self-deception and false belief, cannot understand why self-deception seems so threatening to an agent's integrity.

Interpretation takes into account the rational links between utterances, actions, beliefs, desires, etc., and draws on the assumption that the normative constraints imposed upon mental states are also instantiated in the mind of the agent we interpret. Beliefs – both true and false – play an important role in interpretation since they show how someone takes the world to be. Their part here depends on their fundamentally two-dimensional character: they have normative as well as explanatory roles to play. They are used to explain the behaviour of others and ourselves, but from the first-personal perspective beliefs are usually not used in their explanatory function, but as part of our deliberative stance, and are taken to represent the world as it really is. Indeed, the explanatory value of beliefs depends on the assumption that agents take their own beliefs to be normatively constrained, linked to how the world really is, and thus providing good reasons for acting or forming other beliefs.

Self-deception, by contrast, only enters the picture at a stage where an interpreter has already determined beliefs and their truth-values, and where we simply assume that the belief in question is false. We can imagine charging someone with self-deception, and when we learn some more details about how she came to acquire the belief in question we change our judgement – that she deceives herself – and accept that she was simply mistaken, without necessarily changing our judgement that her belief is false. Self-deception, like belief, can also be said to be a 'normative' concept; but here 'normative' has a different meaning. In attributing self-deception, we make a judgement about the faculties of the agent and the adequacies of her beliefs that is far more damning than any judgement about mistaken beliefs *simpliciter*. And belief is 'normative' in respects in which self-deception is not: beliefs ought to be constrained by the world, by what is true, and there are no similar constraints on self-deception. We have no idea what self-deception 'ought' to be or to achieve.

These considerations proscribe the conclusion that the concept of self-deception must be treated in analogy to that of false belief, and show that the danger of a *reductio* is less imminent that it may have seemed. But we still lack a positive account of the impact these differences make on the process of projective imagination, and thus on the formation of mental concepts on Davidson's account.

8. Projective Imagination and Self-Deception

To proceed further, let us examine carefully the distinction between 'normativity' as applied to beliefs and to self-deception. When we judge that a belief is false, we think it does not truthfully represent the world. This requires an external perspective on the belief that is not open to the agent who holds it. Sometimes agents form false beliefs not because they make a mistake, but because the information they possess is so limited that the best understanding of that information leads to a false belief. Self-deception is an entirely different matter: it requires an internal failure of the agent, a failure that is independent of the input she receives from the world and for which she cannot be exculpated by reference to insufficient information. In self-deception, the agent tries to form a particular belief, whatever the evidence against it; and this requires an intentional manipulation of her own belief which is 'irrational', not just 'false'. The problem does not lie in the access she has to information, but in the use she makes of the information; insofar as the interpreter must assume that the agent is more or less rational, it must *prima facie* be within the scope of the agent's capacity to avoid this problem.

This discussion of the differences between false belief and self-deception clarifies the main difficulty of an interpretationist understanding of self-deception. For Davidson, the very least that is necessary to attribute self-deception to someone are two contradictory beliefs and an intention that explains how the false – and dominant or avowed – belief arose. The particular power of projective imagination, the reason that strongly suggests that it provides a satisfactory understanding of interpretation, rests on the straightforward link between rationality as employed in the first-personal, deliberative perspective, and the interpreted agent's rationality. Yet this also yields the problem: if the interpreter thinks of herself as rational, she cannot simultaneously believe that her beliefs come about in an irrational way, and she must deny that they come about through self-deception rather than rational deliberation. But if it is impossible for her to think that what she holds is a belief and that it is a result of self-deception, then it seems also impossible for her to imagine that someone else's deliberation could lead to such a state of affairs; yet this is required to make sense of the concept of self-deception.

If it were possible for someone to attribute self-deception to herself, and we maintain our present understanding of interpretation in terms of 'taking up the deliberative stance', then self-deception could not be as fundamentally irrational as

Davidson claims. If self-deception – this complex of beliefs and intentions – can be captured in projective imagination, then it would become part of our interpretative practice in the sense that we could straightforwardly attribute to someone the combinations of beliefs and intentions characteristic of self-deception, incorporating the links between them into our picture of the mind in the same way in which we have already incorporated the rational links between beliefs and actions, between beliefs and beliefs, or between utterances and thoughts. This would assimilate the relations between beliefs and intentions involved in self-deception to those relations governed by the demands of rationality, and doing so would undermine the distinction between irrationality and rationality that Davidson sets out to rescue.

There is an alternative way of explaining about how self-deception could be attributed: by actually deceiving oneself, i.e. by forming two contradictory beliefs linked by an intention of the right kind. But although nothing *per se* excludes the possibility that the interpreter sets out to deceive herself, doing so cannot help her understand the concept of self-deception: to deceive herself, the interpreter must remain unaware of the very act of self-deception, hence for projective imagining to succeed she must also remain unaware of her manipulative intention and the contradictory nature of her beliefs. Projectively imagining self-deception in this sense cannot account for our use of the concept, but only shows that self-deception as a phenomenon can apply to everyone.[96]

At the bottom of this difficulty lies the fact that, given Davidson's account of interpretation, we cannot conceive of the different components of self-deception – the contradictory beliefs and the intention – as distinct from each other. We can understand what it is to hold a false belief; we can also understand what it is to hold a true belief. We can even understand why and how someone could come to form the intention to acquire a false belief: if the truth hurts too much, and she wants to avoid this pain, she must avoid the truth (to put it in the form of a Davidsonian practical syllogism). Each component by itself can be part of deliberation and thus projective imagination, even the intention that proved so difficult to fit into the picture of the mind in the previous chapter. If our explicit aim is to reduce psychological pain, we can form the intention of manipulating our own belief

[96] None of these arguments entails that one cannot apply the concept of self-deception to oneself; like other mental concepts it is univocal. In this respect, it does not diverge from the concept of a false belief: one knows it applies to oneself as well as to others, but cannot take one's own belief to be false without simultaneously giving it up. An agent can recognise that her past actions are cases of self-deception, since she can admit that her past belief was false. Also cf. section 6 above.

formation processes. Projectively imagining forming that intention and even acting on it does not lead to any contradiction. The project fails, we argued earlier, because beliefs must be discounted when one becomes aware that they were not formed in response to how the world really is; but its failure does not form part of imagining having and acting on the intention. Yet all this implodes at once when we combine the beliefs and the intention in projective imagination, since then the agent should become aware of the falsehood of her beliefs and is rationally required to abandon the false ones.

If self-deception could be interpreted in the same fashion in which beliefs and other mental states are, we would either have to give up our notion of self-deception as irrational, or reject the view that interpretation is governed by the requirements of rationality that we took to be characteristic for the deliberative stance. The former cannot be an option for us, since it is equivalent to abandoning Davidson's project: explaining how self-deception can be irrational *and* fit into his picture of the mind. The latter would require rejecting the proposal developed in the first part of the chapter; and in the end, it would require rejecting Davidson's idea that the mental vocabulary is committed to the ideal of rationality. The argument could possibly be rejected by claiming that the use of the deliberative stance in interpreting others is unnecessary, but the proponent of this strategy would have to offer another account of interpretation, and the chances of finding an equally compelling account are slim. The proposed account of deliberation and projective imagination is the most adequate interpretation of Davidson's interpretation theory if we take seriously the demands of univocality and Davidson' emphasis on rationality, and we give up this stress on deliberation, we face the difficulties of explaining the univocality of mental concepts that this account successfully overcomes.

9. Projective Imagination and Mental Boundary

The discussion up to now might seem insufficiently charitable, as it leaves out a central feature of Davidson's account of self-deception: the mental boundary. This 'conceptual tool' was explicitly introduced to facilitate the attribution of contradictory beliefs required for self-deception. But we can now see that in addition to the 'psychological' difficulties discussed in the last chapter (especially sections 5 and 6), there are also 'systematic' difficulties with this notion. The problem concerns the relationship between the mental boundary and the process of

interpretation. Either the mental boundary forms part of our concept of self-deception and only enters the picture when events have been interpreted, leaving no space for contradictory mental states; or the mental boundary forms part of the very foundations of the mind and affects how we interpret someone from the very beginning, but undermines the distinction between rationality and irrationality.

The former option is in line with the thought that self-deception is partly an explanatory concept that elucidates how the agent comes to hold the particular combination of (contradictory) beliefs ascribed to her prior to the judgement that this is an instance of self-deception. But this raises the question how we could ascribe this set of beliefs if the aim of every interpretation is to rationalise the other person's behaviour. By the time we allow the mental boundary to enter the picture, events have been given an interpretation, and form part of a network governed by the ideal of rationality. Davidson has not made clear how we could ascribe the contradictory beliefs to an agent without first introducing the mental boundary into the process of interpretation. Hence, introducing the 'conceptual tool' after the interpretation does not help Davidson to solve his problem: by the time the 'mental boundary' is meant to do its work, no work is left to be done.

If, on the other hand, the notion of a mental boundary is already employed in our interpretative practice, we face a serious dilemma: either we insist – as we must if we take seriously Davidson's central idea – that interpretation is 'rationalising', and the mental boundary itself becomes part of our understanding of rationality. This undermines the irrational character of the mental boundary as well as that of self-deception, given that Davidson locates the irrationality precisely where the mental boundary enters the picture. Or we emphasise the irrational nature of the mental boundary, in line with the thought that for someone to be rational she must knows what she thinks, or can at least know this after some deliberation. Since the mental boundary is the device that is intended to overcome this 'transparency' in our picture of the mind, it seems to be required by Davidson's account of self-deception that we insist on its irrationality. So even if we might in theory have the option of holding contradictory beliefs, in practice this ought not to happen if interpretation is governed by the ideal of rationality. Thus, if Davidson's 'mental boundary' is to fulfil the role Davidson assigns to it, it can neither enter at the stage of interpretation nor afterwards. It comes always either too early or too late to help us make sense of self-deception.

10. Conclusion

The particular role the notion of self-deception plays for us cannot be accommodated in Davidson's picture of the mind. On the proposed reading, projective imagination is the road to understanding the use of third-personal mental concepts, and it lets us down when it comes to self-deception. Thus, there is a fundamental shortcoming in Davidson's theory: we cannot use his account to make sense of the possibility of thinking of ourselves and others as irrational. Yet self-deception and other forms of irrationality form part of our psychological vernacular, and Davidson claims to be able to account for them. The problem of locating self-deception in his picture of our mental life in between the fully rational and the purely causal points to a blind spot in his philosophy that we will analyse in detail in the last chapter.

Chapter Four: Re-Interpreting Interpretation

In this chapter I want to draw together the threads of the argument, and offer a conception of the mental realm that accounts for self-deception while keeping in place most of Davidson's insights on the nature of the mind and the relations between first- and third-personal perspectives. First, I argue that we can trace the problems we already identified to an assumption about the nature of interpretation that underlies many of Davidson's arguments, or at least many readings of these. I then propose an alternative understanding of interpretation, and set out to show that it enables us to make sense of the concept of self-deception without succumbing to the problems we located. The picture will be recognisably similar to Davidson's account of self-deception and the mind; but it clearly distinguishes the different levels of analysis which are required for a proper explanation of self-deception and other irrational phenomena, and thus avoids the difficulties Davidson's account encounters.

1. The Underlying Problem

In the previous chapters we assessed Davidson's account of self-deception, first as a self-contained argument, then in the broader context of his philosophy of mind and psychology. We discovered that the account is unpalatable as it stands: when we get to the bottom of it, all we find are purely causal interactions between events, which can only be explained in physical terms and thus fail to fit our bill, since ordinary talk about self-deception is lodged at a personal rather than a subpersonal level. This reveals a problem: Davidson's account of the mental cannot accommodate notions that are psychological without being fully rational, like self-deception.

The underlying problem is that on Davidson's account irrational events occur when the non-mental impinges upon the mental from the outside. If we think that self-deception and irrationality are as much part of the mind as rational beliefs and behaviour, Davidson's account forces us to adopt a picture in which some psychological phenomena are only derivatively mental: what we identify as 'irrational beliefs' must first and foremost be described and explained in physical, and only secondarily in mental terms. Our account ought to reflect that, although we

can think of ourselves in such alienated fashion, it is not the ordinary way we relate to ourselves even when we recognise that we act irrationally.

This is a shortcoming that Davidson himself ought to recognise: his treatment of the notion of rationality and its ambiguities implies that Davidson must avoid creating a divide between fully rational behaviour (describable in mental terms), and less than fully rational behaviour (explainable only in physical terms). Rationality can be negated in two different ways: either we oppose rationality to irrationality, or we oppose it to arationality – either the requirements of rationality are violated, or they simply do not apply. Davidson recognises such a distinction when he insists that only those things in the world that can ordinarily count as rational – persons, beliefs etc. – can also be irrational, whereas the rest of the universe – the paradigmatic rocks, everything that is 'in itself' rather than 'for itself' in Sartrean terminology – is arational.[97] Arational things are not subject to the demands of rationality and can neither violate nor satisfy them. Irrationality and its negation *together* characterise those things that are subject to the 'requirements of rationality'. Thus, we must distinguish between the set of objects to which rationality requirements apply, and its subset, formed by the objects which fully conform to these requirements. Since self-deception is 'irrational', she who deceives herself is subject to the demands of rationality though her actions do not comply with those standards. The realm of the mental is then to be characterised not by its actual compliance with these requirements, but by its being subject to them. Persons and animals with minds (if they exist) do not necessarily fulfil all rationality requirements, but we assess them under the particular perspective of rationality.

Davidson's account forces open a gap between the irrational and the mental proper. He overcomes the traditional dualism of the mental and the physical by arguing that they are not two different types of substance, but two different ways of describing events; thus it is unproblematic that some events can be both mental and physical, since it not particularly difficult to think that the same event can be talked about in more than one way. As long as the respective vocabularies incur different commitments, these descriptions do not come into conflict. But the dualism Davidson exorcises at the front door enters again at the back: as our mental vocabulary is committed to rationality, what is not rational cannot be captured in

this vocabulary, and to talk about, explain, and understand it, we must make use of physical terms. Davidson's account of the mental starts from an important insight: that we cannot explain the mental in terms of physical laws because our mental vocabulary is committed to the ideal of rationality. But in his eagerness to emphasise what keeps apart the two domains, Davidson also creates the situation in which we now find ourselves: it becomes difficult to talk about human beings as psychological subjects when their behaviour is less than fully rational.

Our challenge is to find a way of talking that does not cut loose the mental from either its commitment to rationality or its links with the physical, yet makes it possible to talk about violations of rationality without falling into the abyss where only physical descriptions are adequate. We must come up with an alternative picture that overcomes the choice between 'mental short circuit' and 'complete rationality' forced upon us by Davidson's explanation of self-deception; we must close the conceptual void that emerges in Davidson's account of the mind when irrationality enters if we want to accommodate our practice of blaming people and holding them responsible for self-deception. As Davidson's account stands, we can only explain instances of irrationality by citing events under their physical description. But the blame game we play involves talking about beliefs, desires, intentions, and other notions that are linked to the idea of a person rather than a being whose behaviour we explain by referring to the laws of physics, chemistry, and physiology. We cannot blame people for what occurs in their stomach, so we cannot hold them responsible for what happens in their brain when the physical description under which the events are assessed cannot distinguish putatively mental from digestive events.

2. The Source of the Difficulties

We can trace these problems to a widely shared assumption about the nature of interpretation, which I will show is by no means necessary for an adequate understanding of interpretation theory. In a theory as sophisticated as Davidson's, there is more than one way to trace the emergence of a problem, since it is usually a combination of different assumptions which leads to difficulties; so focusing on the notion of interpretation is only one way to solve our puzzle. But this strategy reveals in relatively straightforward fashion that certain assumptions that have

[97] See, e.g., PI, p.289: "...the irrational is not merely the nonrational, which lies outside the ambit of

created the aforementioned problems can be discarded without undermining Davidson's overall project.

For Davidson, interpretation requires we see someone's behaviour as governed by rationality. Given Davidson's holism, the assignment of meaning to linguistic and non-linguistic behaviour depends on taking in a whole range of information and using it to make sense of an agent's behaviour. I suggest that the prevalent understanding of the holistic constraint on interpretation is in fact mistaken, and leads to the difficulties that we have identified in Davidson's account. On this understanding, interpretation aims at formulating theories of the agent which cover *all* her actions, past and present, at once. This 'global' view of interpretation, with its emphasis on the globalising or all-embracing character of each act of interpretation, leads to serious difficulties.

The first problem is that it is not clear how the connections between the beliefs in question can be of the appropriate mental rather than purely physical kind if self-deception is attributed as a result of global interpretation. In order to bring out the compatibility of rationalising interpretation and the existence of irrationality, Davidson emphasises that successful interpretation demands the greatest possible global consistency, i.e. over all instances of the speaker's behaviour, and claims that this requirement can sometimes only be fulfilled at the expense of local consistency. If global consistency requires that in a situation we attribute certain beliefs to the speaker, it is always possible that between some of these no rationally explicable links exist. Davidson emphasises that the most important criterion for adequate interpretation is the global coherence and consistency of beliefs; but he also states that achieving this can involve the local assignment of beliefs that are not rationally linked.[98] These beliefs fall short of complete rationality in spite of being viewed from within the rationalising perspective.

But there is a problem, covered up by the innocent expression 'not rationally linked'. Although Davidson's explanation moves in the right direction, it effectively accomplishes precisely the opposite of what we need to explain self-deception. The phenomenon depends on a certain form of 'local' rationality, because the manipulation of one's own beliefs is intentional and thus performed with a particular aim in mind, viz. avoiding an unpleasant (belief-) state. We lack an explanation how this (at least instrumentally rational) act of self-manipulation can

---

the rational; irrationality is a failure within the house of reason."

co-exist with the overall demand for coherence and world-guidedness in beliefs. On Davidson's picture, the interpretative demand for global consistency might force us to simply put next to each other, or see as causally linked, rationally unrelated beliefs. Yet this cannot explain the phenomenon of self-deception: we cannot distinguish self-deception from other instances of irrational belief formation if the emergence of the false belief is explained in purely causal terms, without reference to intentions and hence some limited form of rationality. As we have seen in Ch.2, the belief not-p, though it contradicts the initial belief p, comes about as a result of the agent's holding belief p. The current formulation ignores this connection. Thus, it cannot help us overcome our fundamental problem: how we can simultaneously hold that mental descriptions of events employ a vocabulary committed to rationality and that explain self-deception without resorting to physical descriptions.

The second major problem for Davidson is that we cannot explain a concept like self-deception, which requires attributing contradictory and irrational beliefs to the agent, if mental concepts are constituted by interpretation requiring the fullest possible rationalisation of the agent's behaviour. On the best account so far, interpretation amounts to employing our own rational capacities in imaginative projection; but we also found that we cannot attribute to ourselves current self-deception about a particular issue because when we realise that we hold an unwarranted belief due to considerations of pleasure rather than truth, we discard that belief. Since we understand others by employing the same capacities for deliberation and reflection, we cannot explain how self-deception can be attributed to them either.

Projective imagination employs ordinary means of deliberation, but it is obviously still only projection – the interpreter does not actually deliberate for the interpreted agent, but only deliberates as if she were in the agent's position. Since we cannot trace differences between the interpreter's own beliefs and the beliefs she attributes to others to variances in reasoning, they must be due to differential inputs: others hold beliefs which differ from the interpreter's because they are in a different situation or position. But if we assume that the limits on the information the interpreter draws on in projective imagination are all imposed by the limits of the other person's knowledge – she cannot see certain things the interpreter can see, or has not been told certain things the interpreter has been told – then all differences

[98] PI, pp.301ff.

are due to ignorance. Yet self-deception is different from cases in which someone holds false beliefs out of ignorance. Thus, globalising interpretation makes demands on the interpreter which prevent her from attributing self-deception.

3. Alternatives for Interpretation

We can overcome these difficulties by reconsidering the notion of interpretation, and how it affects our conception of the mental. We have identified as one culprit for our problems the assumption that interpretation must always aim at global consistency, sometimes at the expense of local rationalisation; so we should consider an alternative understanding of interpretation.

Jeffrey Malpas introduces one such alternative in his defence of Davidson's philosophy of language against Michael Dummett's criticism. Dummett argues that Davidson's account of interpretation is overly demanding: "...when we try to take seriously the idea that the references of all names and predicates of the language are simultaneously determined together, it becomes plain that we are thereby attributing to a speaker a task that goes quite beyond human capacities."[99] For Dummett, interpretation demands that interpreters simultaneously take into account all the sentences of a language, and all their knowledge of the world, and fit them together in the most satisfying fashion; and this simply goes beyond anyone's intellectual capacities. Malpas proposes a reading of Davidson that avoids this problem; his proposal will help us resolve the difficulties in Davidson's account of the mind. Malpas' response to Dummett is to "take the injunction to construct 'total' theories as a requirement that theory construction should always attempt to take account of as much of the interpretative evidence as possible. ... Of course, what counts as part of the body of relevant evidence will itself be determined by the interpretative project in which we are engaged."[100]

In fact, two proposals are implicit in this explanation of interpretation. First, what counts as relevant evidence for interpretation is limited, and not everything and everyone must be taken into consideration at all times. Second, interpretation, though ubiquitous, is not continuous. Although we must always interpret, not every act of interpretation is seamlessly linked to all others. The activity of interpreting is guided by our interests: interests in understanding what someone means when she utters certain sentences, or what she does when she moves wooden things on a

[99] Dummett (1975), p.29; quoted in Malpas (1992), p.112

black-and-white chequered board. "Interpretation is an activity which proceeds within localised boundaries and with respect to often fairly narrow interests. The sorts of theories that can be constructed are always only 'partial' or 'localised' theories. They are theories which describe only some portion of the psychological, rather than the psychological as a whole, and which typically operates within some particular framework or context."[101] So interpretation is based on 'occasion-specific theories' rather than on theories aiming at the global explanation of all behaviour past and present of the agent. Irrationalities could consequently be ascribed if beliefs attributed on different occasions were contradictory.

We must immediately obviate a natural objection: Malpas' proposal is incoherent, one might think, because as soon as two beliefs were revealed to be contradictory, the interpreter would have to formulate an overarching theory that incorporates the beliefs attributed on the basis of several interactions while qualifying them such as to discard the apparent contradiction. But this argument just assumes what the proposal doubts: that the overall aim of interpretation is to formulate global theories. The new proposal is committed to the weaker claim that theories employed on different occasions are not completely independent of each other, since they are all used to explain the behaviour of persons, are informed by a general interest in understanding others, and draw on the same (the interpreter's) understanding of rationality. The unity of these theories is inherent in their sharing a source, their dependence on a common interest, and their drawing on the same conception of rationality; it is not necessary to create additional unity by formulating ever more overarching theories and interpretations.

Although this amounts to a reconfiguration of how we think rationality is applied in interpretation, it leaves untouched our ordinary understanding of rationality. Within each project, all our intuitions about rationality are satisfied. We could even think of rationality as applicable to the global assessment of our lives if we set ourselves the project of achieving overall consistency in our attitudes. But what has changed is what we take to be the standard application of rationality: it is not only about achieving consistency across an entire life. Such a project is only one of many in which rationality is instantiated – one particularly dear to philosophers, which

---

[100] Malpas (1992), p.111

[101] Ibid., p.113. Importantly, Malpas thinks that his proposal does not amount to a full-scale revision of Davidson's theory, but only to an unveiling of what is already implicit in Davidson's theory, in particularly as developed in NDE.

may explain the prominence it is accorded by Davidson (or his interpreters), but nonetheless no closer to 'real' rationality than others.

4. The Notion of Relevance in Interpretation

To understand the implications of the new proposal, we must clarify the notion of 'relevance' it introduces. It differs from the ordinary use of the term in the same way in which Davidson's use of 'interpretation' diverges from the vernacular use. While talk about 'relevance' ordinarily assumes that we can assess whether something is relevant or not and offer reasons for this judgement, 'relevance' as it is employed here is not open to such explicit assessments, because it is the starting point of any act of deliberation, reflection, assessment, or interpretation. It is only once a piece of information has passed the threshold of 'relevance' that it enters our thinking about a particular subject matter, and only then can we ask whether it is relevant (or irrelevant) in the ordinary sense of the word. The agent does not reject irrelevant evidence – in the sense that matters here – after pondering whether it is relevant or not for the case at hand, because if this were what 'disregarding irrelevant evidence' required, we would never even start weighing up the evidence. The limits of relevance are thus internal to our thinking. If the boundaries of deliberation were fixed by prior deliberation about what is relevant for the subject-matter, we would wind up in an infinite regress that would make any deliberation impossible. Drawing the boundaries around relevant evidence cannot be a matter of deliberation, but must involve something much closer to not even raising the question whether it is relevant for most of our knowledge. Consequently, taking some things to be relevant for the issue at hand, and ignoring others, cannot be an intentional action: intentional actions require reasoning, and here we are considering the limits of reasoning, which themselves cannot be determined by a prior act of reasoning 'from the outside'.

This argument reflects a thread running through Davidson's theory, according to which there is no outside perspective from which we can determine whether something deserves rationalising interpretation.[102] Interpretation is subject to 'interpretative closure': we cannot assess the success or even the need for interpretation from outside the realm of interpretation altogether, independently of an alternative interpretation. Indeed, this cannot come as a surprise for anyone who

[102] This is one way of reading Davidson's argument in VIC.

thinks carefully about Davidson's 'anomalous monism': since it is impossible to identify law-like relationships between events described in mental terms and events described in physical terms, the process of understanding something as a mental or psychological event cannot start from a prior description of the same event in physical terms. As events are only understood or explained under a particular description, and a description in mental terms cannot proceed from a description in another vocabulary, understanding psychological events requires that from the very start one conceives of them as partaking in the realm of the mental.[103] For Davidson, the interpreter cannot stand back from the interpretative process, because in every act of interpretation she must implicitly judge what is relevant and what is not relevant for understanding others.

5. Interpretation and Evidence

The notion of 'relevance' helps us explicate the differences between first- and third-personal perspective that concerned us earlier. When an agent thinks that a piece of (accessible) evidence is relevant for the subject-matter she is concerned with, she will take it (for that reason) into account. From within the first-personal perspective, the agent must always believe that she has taken into account all the relevant evidence available to her; if she realises that she has not drawn on some relevant knowledge she possesses, she must reconsider her belief in the light of the new relevant information.

The situation is different for the interpreter who takes a third-personal perspective on the agent's beliefs: she is in a position to see (or think she sees) that some relevant evidence is being ignored. In projectively imagining an agent's mental states, we must set aside some evidence in order to mirror the situation in which the agent finds herself. In contrast, a first-personal deliberator who aims at the best possible assessment of a situation, cannot ignore any beliefs and must take into account all the relevant evidence without any qualifications. The distinction between taking into account all the evidence that is relevant to a particular subject matter, and taking into account all the evidence than can be known *and* is seen as relevant from a particular perspective – that of the person being interpreted – is required to make sense of the possibility of imaginative projection and the thought that interpretation is similar to, but not the same as, first-personal deliberation.

[103] This provides additional fuel for criticism of the unlikely couple Dennett (1971) and Mulhall

By introducing the distinction between relevant and irrelevant evidence, we can improve on the problematic view that differences between the interpreter's own beliefs and those of the agent she interprets must be traced to differential access to information. On that view, everything there is to know about the world has to be taken into account to interpret others, making it almost impossible to explain how the interpreted agent could be attributed ignorance about something that the interpreter knows. This motivates the introduction of a 'mental boundary' as an external tool that explains how some knowledge the interpreter possesses is sometimes not available to the interpreted subject. On the localised account of interpretation, such an artificial and external barrier becomes unnecessary; instead, it becomes part of the very nature of interpretation that certain things are outside of what we take into account. As a result, discrepancies between the interpreter's beliefs and those of others are not always traced to divergent information; instead, such discrepancies can be explained by different assessments of the relevancy of some data for a particular belief.

Just as someone who forms a false belief due to incomplete information counts as mistaken, so we must allow that not all assessments of the relevancy of evidence are equally valid or good. The interpreter must take her own assessment of the situation to be correct, and can thus reasonably blame others for their mistaken views on what is relevant for a certain subject-matter. Davidson explicitly says that when we reach a fundamental aspect of rationality, "the distinction between the standards of rationality of the agent himself and of his critic merge"[104]; but although he usually focuses on standards of deductive rationality, such as the principles of decision theory, the current proposal makes explicit that the same demands apply to non-deductive rationality. If someone fails to reason in accordance with the basic rules of decision theory, she violates her own principles insofar as any interpreter must assume that she accepts these principles if her behaviour is to count as meaningful and interpretable. Similarly, if someone fails to realise the relevance of certain pieces of information that the interpreter sees as fundamentally relevant to the subject-matter, the interpreter must take her to violate her own standards.

The discussion of 'relevancy' brings to the fore again a problem we raised earlier: for Davidson, when we assess whether someone is deceiving herself about p only those things count as evidence for p or not-p which the self-deceiver counts as

(1990), who for very opposite reasons think that interpretation theory is 'instrumentalist'.

such.[105] Yet it is difficult to see how the interpreter could know what someone else counts as evidence except by realising that the other person employs it in forming a belief; and this is precisely what someone who deceives herself fails to do. But we must also recognise the valid motivation behind Davidson's claim: because self-deception is a matter of internal incoherence, we cannot simply blame someone for failing to take into account evidence that she could not even become aware of. The localised view of interpretation enables us to respond to Davidson's worry as well as to our earlier criticism: it is correct that in self-deception we do not blame someone for mere ignorance; hence only facts of which the agent could have become aware, i.e. which she is not barred from accessing, are relevant for assessing whether she deceives herself. But this does not entail, as Davidson seems to think, that only those facts which the agent thinks of as evidence for a belief are relevant. Among those facts of which the agent could potentially be aware we must further distinguish between those which she takes to be relevant and those she takes to be irrelevant for forming a certain belief. This will help us explain how someone can know all the relevant facts, yet fail to draw the right conclusion, and why we call someone like that 'irrational'.

This more subtle delineation of the relationship between the activities of the interpreter and the activities of the interpreted subject helps us understand how we can attribute self-deception to someone even if interpretation proceeds by imaginative projection. By drawing on the differences in assessments of relevancy between people, the localised understanding of interpretation can make sense of our capacity to adopt different perspectives and to imaginatively project ourselves into the standpoint of others we seek to understand.

6. Projection, Relevance, and Self-Deception

This explanation of the interaction between interpreter and interpreted subject has important consequences for the possibility of attributing contradictory beliefs. This, remember, is our problem: from the perspective of the interpreter, attributing self-deception to someone requires recognising (a) that the person holds a false belief, and (b) that she holds it as a result of an intentional action of some sort, and thus due to a desire to hold this (false) belief. But this requires attributing mental states or beliefs the interpreter could not consistently hold herself; since interpreting

[104] II, p.346

others is a matter of employing her very own understanding of rationality, it becomes difficult to see how the interpreter can attribute to someone combinations of beliefs which she herself would not hold.

The notion of relevance now provides an additional explanation for the divergence between the interpreter's view of the world and that of others: their assessments of what is relevant can differ. Thus, the interpreter employs her own understanding of rationality within particular projects, but she can also permit that others draw different boundaries and fail to take into account relevant evidence. Most importantly, the interpreter is not required to think of either her own interpretation or the other person's deliberation as continuous: she can attribute to someone a certain belief on the basis of linguistic and non-linguistic behaviour on one occasion, but also attribute a contradictory belief, or an intention to form a contradictory belief, on the basis of other evidence. She does not need to discard either of these, because she can make sense of the idea that someone may fail to put together intentions or beliefs, since she knows that mental activities are 'occasion-specific' and limited by considerations of relevance.

The interpreter rationalises the self-deceiver's behaviour within the limits of particular projects, but not across all of them at once. This allows her to attribute the (intentional) self-manipulative act, the true belief p, and the false belief not-p without requiring that the self-deceiver puts them together. If the self-deceiver draws the boundaries of relevance so that her intentions do not figure in her assessment of the value of the evidence for not-p, then she can simultaneously hold the contradictory beliefs and intentions. Interpreters can attribute the different beliefs to her, each of which is rational within its limited context, because (interpretation reflects that) an interpreted agent need not always take into account all the evidence that (the interpreter thinks) is relevant.

Self-deception can be attributed because this ascription of beliefs and intentions does not collapse when assessed from a third-personal perspective. While first-personal deliberation requires that one disavows all false beliefs and draws the boundaries of relevance according to what seems best, the third-personal view makes less stringent requirements: in projective imagination the limits of relevance are not drawn according to the interpreter's best understanding, but follow (what the interpreter takes to be) the interpreted subject's understanding of the case – even if

[105] Cf. Ch.2, section 3.

that is more limited. For example, we can think of the formation of an intention to acquire the belief not-p as a project governed by the desire to limit the pain that the belief p (whether true or false) causes. An interpreter can distinguish between different interpretative projects pursued by the interpreted subject, and interpret the actions, beliefs, and intentions in each case according to the subject's understanding of each particular project. When the interpreter compares the results of her interpretation of the interpreted subject's actions over time, she can therefore hold on to earlier interpretations that now seem less than fully consistent. The interpreter can keep in mind that from within a limited perspective the intention is rational, yet from a different – and more important – perspective it is irrational.

This explanation is substantially different from the one we reject at the start of this chapter, according to which self-deception is attributed when global interpretation yields an assignment of two contradictory beliefs. That explanation made the assignment of contradictory beliefs seem fortuitous because it failed to account for the specific feature of self-deception, the intentional action that linked the two beliefs. The new proposal avoids this problem because it enables us to attribute both the contradictory beliefs and the intention that leads from one belief to the other: the intention to form a false belief can be rationalised, and, most importantly, that rationalisation can be kept in place even when the 'bigger picture' is considered.

This explanation also makes clear why we cannot ascribe current self-deception about a particular subject-matter S to ourselves, yet can attribute it to others and our past selves. We have up to now emphasised that the deliberative stance plays a particular role in our thinking about the mental life of persons. But we can now see that we do not see ourselves exclusively through the lens of deliberation; we can also adopt a third-personal perspective by taking our own mental states as irreversible data which can be assessed and judged as to how they fit into a wider pattern of other mental states and behaviour. This requires distancing oneself from the beliefs one assesses in this way, which exposes a fundamental difference between concepts like belief, desire, etc., and self-deception: if the agent judges that she is deceiving herself, she must take a third-personal perspective onto herself, and think of herself not as the fully rational agent with whom she usually identifies, but as a fallible being struggling to live up to the ideal of rationality.[106]

[106] See Gilbert (1971) for a related discussion of the self-ascription of vices.

7. Lazar's Criticism Revisited

We have just sketched the way in which the revised understanding of interpretation can avoid the problems that arose when we tried to find a place for the concept of self-deception in an interpretation account of the mind. The viability of the claim that the interpreter can interpret the self-deceiver's actions *and* see that they are irrational depends on how much sense interpreters can really make of the idea that an agent can deceive herself. Answering this requires reconsidering Lazar's criticism of Davidson's account of self-deception. If someone deceives herself, it is not enough for her to want to believe that not-p is true; she must also come to believe that there is better evidence for not-p than p. What sets self-deception apart from other instances where agents hold contradictory beliefs is the source of one of the beliefs: evidence for the false belief is acquired by intentional self-manipulation. The difficulty is to explain how someone can simultaneously believe not-p, and know that she only acquired this belief as a result of intentional self-manipulation. We argued that the intention to form the belief not-p only counts as successful if its content is linked to the belief not-p, and thereby showed that Davidson's mental boundary could not overcome the impairing effects the intention has on the formation of the false belief.

How can the revised account accommodate this powerful criticism? In the previous section we proposed that the self-deceiver forms the intention and the two beliefs as part of distinct projects, which she need not bring together. It would seem, at first, that this picture also falls prey to Lazar's argument: if the self-deceiver forms the initial belief p and the intention to believe not-p, and finally acts so as to acquire evidence for not-p, then her knowledge of the intention to acquire this evidence must make it clear to her that the belief founded on this evidence is not reliable and must be disavowed. But this argument does not do as much damage here as it did against Davidson. The problem for Davidson's proposal is that the mental boundary completely holds apart the contradictory beliefs, or the intention and the belief it is about. The revised proposal does not require anything as strong as that, because it can make sense of situations in which the agent has access to some information, but does not consider it for her project because it does not seem relevant. Thus, the agent can be aware of intention and action leading to the acquisition of a belief, yet not take it into account for her formation of the belief not-p due to its apparent irrelevance.

Consider a situation in which a critic points out to the self-deceiver that the evidence she possesses for her belief is skewed. Unless she discounts her belief, the self-deceiver must somehow deny the cogency of that claim, and to do so she must deny either that what the critic says is true, or that it is relevant. If the critic reminds the self-deceiver that she only attended to information favourable to her preferred position, and neglected checking whether the sources cited therein where reliable, the response can be either 'That is not true; I read some other stuff too, but it was not very convincing' or 'Perhaps, but why does that matter?' (That is, 'Why do you bother me with so obviously irrelevant stuff?') If the response is the former, we can interpret the agent in two ways. First, the self-deceiver could be positively blind to what the critic points out, i.e. she could simply not remember her own actions which lead to the acquisition of the evidence. But this must be explained either by an intention to forget, which would bring us back to where we were before, or by physiological shortcomings, which would leave us in the dark about the basis for the moral excoriation that is a common response to self-deception: physiologically induced blindness is hardly something for which we can blame agents. Second, she could see what the critic points out, yet claim that it is not true; but if this denial leads to her actual believing what she says, we have an instance of successful self-deception. This would again require explanation, and we have not made any progress.

Instead, the self-deceiver must fail to see the relevance of the information the critic provides. The distorted evidence is used because its self-manipulative origin is not considered. Successful reasoning depends on taking into account all the relevant evidence and information; whether an agent is judged to reason successfully and is counted as rational thus partly depends on the boundaries of relevance the interpreter draws around a particular project.

8. Distinguishing Levels of Explanation

Thinking of interpretation in this localised fashion provides the means for thinking of self-deception as an intentional and thus partly rationalisable, yet overall irrational phenomenon. Self-deception can count as intentional because we can identify an island of rationality in this otherwise irrational phenomenon: for example, it makes sense to minimise suffering, and if we simply focus on the fact that acting in a particular fashion – manipulating our evidence acquisition and

holding a false belief – alleviates pain, we can see contours of rationality. But this by no means entails that the interpreter has rationalised something inherently irrational, because if she thinks about the issue in a broader context, i.e. enlarges the frame within which she looks at the action and takes into account additional evidence which she considers to be relevant, the irrational character of the phenomenon comes again to the fore.

That someone fails to recognise contradictions in her belief set, or to realise that her beliefs are the upshots of self-manipulation rather than warranted by the way the world is, requires further analysis. It cannot be *inherently* irrational to not consider everything we have access to, because this would reinstate the overextended demands of holism; but in cases of self-deception it certainly cannot be rational to ignore relevant evidence either. Nor can we explain that someone ignores the relevance of the self-manipulative intention by a further intention not to take the initial intention into account, as this would simply reintroduce the problem: why does she not put her new intention together with the mistaken belief?

To solve the problem, we must draw a distinction between the level at which such questions or explanations have a place, and a level where we cannot make sense of them. Throughout the discussion we have insisted that Davidson fails to account for the particular link between the contradictory beliefs when he introduces the mental boundary that keeps them apart. In addition, Davidson does not explain how we could suddenly stop deliberating and erect a mental boundary. The latter problem does not arise if the shortcoming of the agent does not consist in disregarding evidence she knows is relevant, but instead in ignoring the relevancy of information she possesses. Davidson's problem initially arises because the emergence of a mental boundary that prevents the agent from going through with deliberation cannot be explained, yet urgently requires explanation. But on our account, the boundary does not unnaturally prevent the agent from deliberating; rather, the agent naturally stops deliberating where she assumes she has taken into account all the relevant evidence – her stopping here reflects her conviction that no further investigation is required. Explaining why someone stops here rather than there in collating evidence is fundamentally different from the intention-invoking explanations usually proposed for self-deception. From within the deliberative perspective, the self-deceiver simply insists that there is nothing further to be taken into account, that further details would not change her belief, etc. From the outside,

the interpreter draws the boundaries differently, and blames the agent for ignoring relevant evidence. It is from this perspective that it becomes clear what has gone wrong with the self-deceiver: she has failed to take into account what is (from the interpreter's viewpoint) obviously relevant, viz. the role of her own intentions for assessing the reliability of evidence for a particular belief. This mistake is so fundamental that it undermines her rationality.

We do not need to invoke intentional actions to explain this shortcoming, because the agent does not intentionally disregard the relevance of the information that her evidence for not-p is the result of self-manipulation; indeed, it is something to which the very concept of intentionality is inapplicable insofar as it requires reason and reason cannot be used to fix its own boundaries. From the perspective of the interpreter, the agent who does not see that her intentional self-manipulation is relevant for the assessment of her current belief is wrong; the interpreter thinks her action has lead to the acquisition of skewed evidence and the formation of a false belief, and therefore attributes self-deception to the agent.

Describing the phenomenon of self-deception in this way does not make it devoid of moral relevance; but it shows that we cannot seek to explain its moral import in terms of intentional actions. The moral implications of self-deception are not due to intentions, but to something more fundamental to our thinking about moral agency: its very foundations, which provide the basis for our talk about morally relevant actions. By denying that her intentional actions are relevant for beliefs she subsequently forms, the agent fails to 'know her own actions', for which she bears a special responsibility; hence she is blameworthy.[107]

The important step forward we make here is that we avoid ascribing to the person who deceives herself an incongruent and thus not properly attributable set of beliefs and actions; within the bounds of her reasoning, the agent's behaviour can be rationalised and interpreted – hence we can identify the two contradictory beliefs and the intentional action leading from one to the other. But her inability to recognise that the different deliberative 'projects' bear relevance upon each other enables her to hold a set of beliefs which the interpreter realises are contradictory and inherently irrational.

[107] Cf. Bilgrami (1998), especially part II.

9. Conclusion

The proposed account is only subtly different from Davidson's; but small differences carry a lot of weight in a theory that is as subtle as Davidson's. The solution we offer diverges from Davidson's account of self-deception (that we rejected in the second chapter) because it entails that the limits of reasoning are set by neither intentional action nor physiological causes. That someone who deceives herself can be blamed does not entail that we can explain how she could have failed to take the information about her self-manipulative intention into account; the claim that she disregards relevant information depends on the perspective of the interpreter who draws the boundaries around the domain of relevant information more widely than the self-deceiver.

By thinking of interpretation in terms of projects with limits drawn by considerations of relevance, the new account avoids the contentious notion of a mental boundary. Davidson introduces that notion to explain how an agent could fail to have full access to her own mental states. His starting assumption therefore is that the agent is fully aware of all that the interpreter thinks she ought to be aware of; and this is the very assumption that the new, localised understanding of interpretation can do without. It would be mistaken to think that the localised understanding of interpretation just is global interpretation plus the mental boundary. For Davidson, the mental boundary enters the picture when we realise that agents do not always have full self-knowledge of their own beliefs; there is no further motivation or explanatory role for it. On the new proposal, there are limits to interpretation and deliberation, yet these are not externally imposed but internal to the notion of interpretation and deliberation. They do not require an explanation of the kind Davidson's mental boundary demands, and which he could not offer: he uses the mental boundary as the main device to explain self-deception, but sets it against the background of the globalised account of interpretation, which, as our analysis has shown, cannot accommodate the mental boundary without undermining it.

The fundamental difficulty with Davidson's account is that it does not make clear where we have to draw the line between reasonable explanation and unreasonable reduction of irrational mental phenomena. It fails to make clear how it is possible to explain self-deception without falling back onto hypotheses about the physical foundations of irrational beliefs. This problem can be solved by adopting the more

self-deception is a phenomenon that fits into the mental realm constituted by rationalising interpretation, as Davidson suggests; but it also becomes clear that if we want to make sure that rationalising interpretation does not exclude the possibility of attributing irrational phenomena, we must avoid globalising interpretations. Instead we must trace the natural limits of interpretation that are inherent in our capacity to reason, cannot be explained in physiological terms – they are irreducibly normative in nature, just like the notions of 'rationality' and 'relevance' – and provide the very foundations for our mental lives.

Conclusion

I started this thesis with the aim of elucidating why irrationality, and in particular self-deception, poses such a puzzle for Davidson's philosophy of mind and psychology. After arguing that his own explanation of self-deception is unsatisfactory, I showed that the error, as so often, lies in the detail: Davidson's account of the mental is shaped by a view of interpretation that prevents him from getting a proper picture of irrationality.

The initial analysis of Davidson's theory of self-deception revealed that it could not make sense of the phenomenon without leaving the level of mental descriptions and slipping into talk about physical or physiological causes. To defend the vernacular conception of self-deception, and to make sense of Davidson's view that self-deception involves an intention rather than merely two contradictory beliefs, we had to reject Davidson's explanation of the phenomenon. But from Davidson's account we learned that there are serious difficulties with explanations of irrationality, insofar as such explanations must by their very nature set out to show the phenomenon to be reasonable in at least a limited sense.

Tracing the links between self-deception and Davidson's philosophy of the mind and psychology, we detected more fundamental problems, which are nonetheless related to the ones discovered in the previous discussion. To explain an action or a mental state, we must rationalise it to some extent; but for Davidson, we must view the mental as a whole (constitutively) through the lens of rationality or rationalising interpretation. When we spelled out how we should make sense of such 'constitutive interpretation', we discovered that the very phenomenon of self-deception seemed inadmissible in Davidson's picture of the mind.

Both Davidson's account of self-deception and (especially) his theory of the mind are highly abstract and to a large degree determined by how language reveals the world to be.[108] Our criticism is hence equally abstract: it focuses on the proper ways of describing self-deception and conceiving of the theoretical phenomenon of interpretation. By sketching an account of interpretation according to which it is 'occasion-dependent' and takes into account only what is relevant to the project pursued on that particular occasion, we can make more sense of the idea that

[108] Cf. MTM, pp.199f.

someone could form beliefs and intentions which are irrational when considered together, but can nonetheless co-exist in one agent.

The revision makes two important improvements to Davidson's original proposal. First, we avoid the talk about a 'mental boundary' that we could not make sense of in Davidson's account, and instead make clear that a certain kind of boundary is part of any mental activity, whether rational or irrational. An account that implies that, when drawn wrongly, these boundaries must be explained with reference to physical or physiological causes must also invoke physical or physiological explanations when the agent is rational. If we find the latter unattractive, the same should hold for the former. This revision prevents us from 'falling into the abyss' and having to employ non-mental terms to explain self-deception, which was our main concern in the second chapter. Second, by elucidating the irrational agent's behaviour in terms of 'relevance', we avoid the problem of having to explain the relation between contradictory beliefs and intentions, which was crucial to the demise of Davidson's original proposal.

To the extent that this account can make more sense of the phenomenon of self-deception than does Davidson's, it improves our understanding of the mind and its relation to rationality and irrationality. Insofar as the proposed account retains most of Davidson's theory and explains why many readily accept the mistaken assumptions about the nature of interpretation, our discussion displays the kind of charity that Davidson requires from interpreters.

Abbreviations

ARC	Davidson, Donald (1963). "Actions, Reasons, and Causes". Reprinted in Donald Davidson (2001a), 3-19
WW	Davidson, Donald (1969). "How is Weakness of the Will Possible?". Reprinted in Donald Davidson (2001a), 21-42
ME	Davidson, Donald (1970). "Mental Events". Reprinted in Donald Davidson (2001a), 207-27
AG	Davidson, Donald (1971). "Agency". Reprinted in Donald Davidson (2001a), 43-61
RI	Davidson, Donald (1973). "Radical Interpretation". Reprinted in Donald Davidson (2001b), 125-39
BBM	Davidson, Donald (1974a). "Belief and the Basis of Meaning". Reprinted in Donald Davidson (2001b), 141-54
VIC	Davidson, Donald (1974b). "On the Very Idea of a Conceptual Scheme". Reprinted in Donald Davidson (2001b), 183-98
PP	Davidson, Donald (1974c). "Psychology as Philosophy (with Comments and Replies)". Reprinted in Donald Davidson (2001a), 229-44
TT	Davidson, Donald (1975). "Thought and Talk". Reprinted in Donald Davidson (2001b), 155-70
MTM	Davidson, Donald (1977). "The Method of Truth in Metaphysics". Reprinted in Donald Davidson (2001b), 199-214
IN	Davidson, Donald (1978). "Intending". Reprinted in Donald Davidson (2001a), 83-102
PI	Davidson, Donald (1981). "Paradoxes of Irrationality". In Richard Wollheim & James Hopkins (eds.) (1982), 289-305
RA	Davidson, Donald (1982). "Rational Animals". Reprinted in Donald Davidson (2001c), 95-105
FPA	Davidson, Donald (1984). "First-Person Authority". Reprinted in Donald Davidson (2001c), 3-14
II	Davidson, Donald (1985). "Incoherence and Irrationality". *Dialectica* 39/4, 345-54
DD	Davidson, Donald (1986). "Deception and Division". In Ernest LePore & Brian McLaughlin (eds.) (1985), 138-48

NDE Davidson, Donald (1986). "A Nice Derangement of Epitaphs". In Ernest LePore (ed.) (1986), 433-46

KOM Davidson, Donald (1987). "Knowing One's Own Mind". Reprinted in Donald Davidson (2001c), 15-38

WF Davidson, Donald (1998). "Who is Fooled?". In Jean-Pierre Dupuy (ed.) (1998), 1-18

Bibliography

Belgum, Eunice (1990). *Knowing Better – An Account of Akrasia*. Garland Publishing, New York

Bermúdez, José Luis (2000). "Self-Deception, Intentions, and Contradictory Beliefs". *Analysis* 60/4, 309-19

Beyer, Lawrence (1998). "Keeping Self-Deception in Perspective". In Jean-Pierre Dupuy (ed.) (1998), 87-111

Bilgrami, Akeel (1998). "Self-Knowledge and Resentment". In Crispin Wright et al. (eds.) (1998), 207-41

Cassam, Quassim (ed.) (1994). *Self-Knowledge*. Oxford University Press, Oxford

Davidson, Donald (1963). "Actions, Reasons, and Causes". Reprinted in Donald Davidson (2001a), 3-19

Davidson, Donald (1969). "How is Weakness of the Will Possible?". Reprinted in Donald Davidson (2001a), 21-42

Davidson, Donald (1970). "Mental Events". Reprinted in Donald Davidson (2001a), 207-27

Davidson, Donald (1971). "Agency". Reprinted in Donald Davidson (2001a), 43-61

Davidson, Donald (1973). "Radical Interpretation". Reprinted in Donald Davidson (2001b), 125-39

Davidson, Donald (1974a). "Belief and the Basis of Meaning". Reprinted in Donald Davidson (2001b), 141-54

Davidson, Donald (1974b). "On the Very Idea of a Conceptual Scheme". Reprinted in Donald Davidson (2001b), 183-98

Davidson, Donald (1974c). "Psychology as Philosophy (with Comments and Replies)". Reprinted in Donald Davidson (2001a), 229-44

Davidson, Donald (1975). "Thought and Talk". Reprinted in Donald Davidson (2001b), 155-70

Davidson, Donald (1977). "The Method of Truth in Metaphysics". Reprinted in Donald Davidson (2001b), 199-214

Davidson, Donald (1978). "Intending". Reprinted in Donald Davidson (2001a), 83-102

Davidson, Donald (1981). "Paradoxes of Irrationality". In Richard Wollheim & James Hopkins (eds.) (1982), 289-305

Davidson, Donald (1982). "Rational Animals". Reprinted in Donald Davidson (2001c), 95-105

Davidson, Donald (1984). "First-Person Authority". Reprinted in Donald Davidson (2001c), 3-14

Davidson, Donald (1985). "Incoherence and Irrationality". *Dialectica* 39/4, 345-54

Davidson, Donald (1986). "Deception and Division". In Ernest LePore & Brian McLaughlin (eds.) (1985), 138-48

Davidson, Donald (1986). "A Nice Derangement of Epitaphs". In Ernest LePore (ed.) (1986), 433-46

Davidson, Donald (1987). "Knowing One's Own Mind". Reprinted in Donald Davidson (2001c), 15-38

Davidson, Donald (1998). "Who is Fooled?". In Jean-Pierre Dupuy (ed.) (1998), 1-18

Davidson, Donald (2001a). *Essays on Actions and Events* (2nd ed.). Oxford University Press, Oxford

Davidson, Donald (2001b). *Inquiries into Truth and Interpretation* (2nd ed.). Oxford University Press, Oxford

Davidson, Donald (2001c). *Subjective, Intersubjective, Objective*. Oxford University Press, Oxford

Davies, Martin & Tony Stone (eds.) (1995). *Folk Psychology: The Theory of Mind Debate*. Blackwell, Oxford

Dennett, Daniel C. (1971). "Intentional Systems". *The Journal of Philosophy* 68/4, 87-106

Dennett, Daniel C. (1991). "Real Patterns". *The Journal of Philosophy* 88/1, 27-51

Dummett, Michael (1975). "What is a Theory of Meaning? (I)". Reprinted in Michael Dummett (1993), 1-33

Dummett, Michael (1993). *The Seas of Language*. Clarendon Press, Oxford

Dupuy, Jean-Pierre (ed.) (1998). *Self-Deception and Paradoxes of Rationality*. CSLI Publications, Stanford

Evans, Gareth (1982). *The Varieties of Reference*. Clarendon Press, Oxford

Fingarette, Herbert (1998). "Self-Deception Needs No Explaining". *The Philosophical Quarterly* 48/192, 289-300

Gilbert, Margaret (1971). "Vices and Self-Knowledge". *The Journal of Philosophy* 68/15, 443-53

Hamilton, Andy (2000). "The Authority of Avowals and the Concept of Belief". *European Journal of Philosophy* 8/1, 20-39

Heal, Jane (1995). "Replication and Functionalism". In Martin Davies & Tony Stone (eds.) (1995), 45-59

Heil, John & Alfred Mele (eds.) (1993). *Mental Causation.* Oxford University Press, Oxford

Holton, Richard (2000). "What is the Role of the Self in Self-Deception?". *Proceedings of the Aristotelian Society* 101, 53-69

Ishiguro, Kazuo (1989). *The Remains of the Day.* Faber and Faber, London

Lazar, Ariela (1998). "Division and Deception: Davidson on Being Self-Deceived". In Jean-Pierre Dupuy (ed.) (1998), 19-36

LePore, Ernest & Brian McLaughlin (eds.) (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davidson.* Basil Blackwell, Oxford

LePore, Ernest (ed.) (1986). *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson.* Basil Blackwell, Oxford

Loar, Brian (1981). *Mind and Meaning.* Cambridge University Press, Cambridge

McDowell, John (1985). "Functionalism and Anomalous Monism". Reprinted in John McDowell (1998), 325-40

McDowell, John (1998). *Mind, Value, and Reality.* Harvard University Press, Cambridge (MA)

McLaughlin, Brian & Amélie O. Rorty (eds.) (1988). *Perspectives on Self-Deception.* University of California Press, Berkeley

McLaughlin, Brian (1993). "On Davidson's Response to the Charge of Epiphenomenalism". In John Heil & Alfred Mele (eds.) (1993), 27-40

Malpas, Jeffrey (1992). *Donald Davidson and the Mirror of Meaning: Holism, truth, interpretation.* Cambridge University Press, Cambridge

Martin, Mike W. (1986). *Self-Deception and Morality.* University Press of Kansas, Lawrence

Mele, Alfred (1997). "Real Self-Deception". *Behavioral and Brain Sciences* 20, 91-102

Mele, Alfred (2001). *Self-Deception Unmasked.* Princeton University Press, Princeton

Moran, Richard (1994). "Interpretation Theory and the First-Person". *The Philosophical Quarterly* 44/175, 154-73.

Moran, Richard (2001). *Authority and Estrangement*. Princeton University Press, Princeton

Mulhall, Stephen (1990). *On Being in the World: Wittgenstein & Heidegger on Seeing Aspects*. Routledge, London

Perry, John (1979). "The Problem of the Essential Indexical". *Noûs*, 13/1, 3-21

Ramberg, Bjorn T. (1989). *Donald Davidson's Philosophy of Language*. Basil Blackwell, Oxford

Ryle, Gilbert (1994). "Self-Knowledge". Reprinted in Cassam (ed.) (1994), 19-42

Sanford, David H. (1988). "Self-Deception as Rationalization". In Brian McLaughlin & Amélie O. Rorty (eds.) (1988), 157-70

Sartre, Jean-Paul (1956). *Being and Nothingness*. Translated by Hazel Barnes. Washington Square Press, Philosophical Library, New York

Scott-Kakures, Dion (1996). "Self-Deception and Internal Irrationality". *Philosophy and Phenomenological Research* 56/1, 31-56

Sellars, Wilfrid (1997). *Empiricism and the Philosophy of Mind*. Harvard University Press, Cambridge (MA)

Siegler, Frederick A. (1963). "Self-Deception and Other Deception". *The Journal of Philosophy* 60/22, 759-64

Steffen, Lloyd H. (1986). *Self-Deception and the Common Life*. Peter Lang, New York

Stich, Stephen (1983). *From Folk-Psychology to Cognitive Science: The Case Against Belief*. MIT Press, Cambridge (MA)

Strawson, P.F. (1959). *Individuals: An Essay in Descriptive Metaphysics*. Oxford University Press, Oxford

Van Fraasen, Bas (1988). "The Peculiar Effects of Love and Desire". In Brian McLaughlin & Amélie O. Rorty (eds.) (1988), 123-56

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Blackwells, Oxford

Wright, Crispin, Barry Smith and Cynthia Macdonald (eds.) (1998). *Knowing Our Own Minds*. Clarendon Press, Oxford

Wollheim, Richard & James Hopkins (eds.) (1982). *Philosophical Essays on Freud*. Cambridge University Press, Cambridge