Chapter 8: Further applications*

Keith Frankish

The theory developed in this book is a wide-ranging one, with applications to many other issues in the philosophies of mind and action. If folk psychology does require a two-level framework of the sort outlined, then many existing debates will need re-evaluation. The framework will require new distinctions, permit new explanatory strategies, and cast old problems in a new light. There should also be implications for psychology and cognitive science. In this final chapter I shall briefly consider three areas in which supermind theory may have application: akrasia, self-deception, and first-person authority. In each case my aim will be modest. I shall not attempt to survey the literature or to argue for the superiority of my approach, but confine myself to sketching the application of the theory and briefly indicating how it can resolve some of the puzzles associated with the phenomenon in question. The discussion will also include some remarks on the nature of intention. I shall close the chapter with a section outlining some possible applications of supermind theory in scientific psychology.

1. Akrasia

Akrasia and self-deception are a pair of puzzling phenomena. Though apparently common, they seem to involve the violation of some basic norms of rationality, and it is not obvious how a unified intentional agent could suffer from them. They also present a special challenge to those who take an austere view of the mind. If mental explanation involves a presumption of rationality, as austere theorists maintain, then it is hard to see how attributions of akrasia and self-deceit could be warranted. The only resource for such theorists, it seems, is to adopt what is sometimes called a *partitioning strategy* – that is, to suppose that the human psyche is partitioned into distinct subagents, each internally rational but in conflict with the other (Davidson 1982; Pears 1984; I borrow the term 'partitioning strategy' from Mele 1987a, 1987b). I want to propose an alternative approach. Both Cohen and Dennett suggest that a two-strand theory of belief is needed in order to explain akrasia and self-deception (Cohen 1992, ch. 5; Dennett 1978a, ch. 16), and in what follows I shall show how the version developed here can provide a robust account of them, compatibly with an austere view of the basic mind. I shall deal with akrasia in this section and with self-deception in the next.

---

* This is the author's version of Chapter 8 of Keith Frankish, *Mind and Supermind* (pp. 203-33), Cambridge University Press, 2004. It does not include final publisher proof-corrections or book pagination. Publishing details of works cited in the chapter can be found in the consolidated References section in the book (pp. 235-45). For information on how to obtain the book, see http://www.cambridge.org/gb/knowledge/isbn/item1117184/.

1.1 Akrasia and the supermind

By 'akrasia' I mean acting against one's better judgement. The akratic agent judges that it is better, all things considered, to perform action A rather than action B, and yet, without revising or abandoning that judgement, freely and intentionally performs B. Note that in saying that the akratic agent judges it *better* to perform action A, I do not mean that they judge it *morally* better – just that they judge it better for *them*, given their needs, desires, and interests. Thus a criminal might judge it better to kill a witness rather than let them escape, and would be akratic if they failed to act upon this judgement. In other words, akrasia need not involve a failure to do what is right, merely a failure to do what one's practical reasoning tells one to do. Note, too, that akrasia is different from inconstancy in judgement. A person who keeps revising their judgements of what it is best to do is inconstant, but not akratic.

I think we all recognize akrasia as a condition from which we occasionally suffer. Yet it is puzzling. Why does the agent perform the akratic action, B, given that they judge it better to perform the other one, A? If the akratic action is intentional, then it must be done for a reason (or so a very common view has it). Yet the akratic agent seems to have no reason for performing B. Any reason they might have had has already been outweighed by other considerations favouring A. It seems that the agent both has a reason for performing B and has none – that their action is both intentional and not intentional. (I am using 'reason' here in an internalist sense, of course.) Another problem, highlighted by Donald Davidson, is how to reconcile the possibility of akrasia with two plausible principles linking judgement with desire and desire with action. If an agent judges that it would be better to perform action A rather than action B, then they want to perform A more than B; and if an agent wants to perform action A more than action B, and believes they are free to perform either, then they will intentionally perform A if they intentionally perform either. So how can people intentionally act against their better judgement? (See Davidson 1969; the principles mentioned are the ones Davidson calls *P2* and *P1*, in that order.)

There are many ways of responding to these problems (for a survey, see Walker 1989). Some writers see akrasia as resulting from the application of incommensurable values (Wiggins 1980); some argue that it involves the 'usurpation' of behavioural control by an unruly desire (Pears 1982); while others endorse a partitioning strategy, positing 'two semiautonomous departments of the mind', one favouring the judgement, the other prompting the akratic action (Davidson 1982, p. 300). No consensus has emerged, however, and I want to suggest a new approach, grounded in supermind theory.

A central feature of akrasia is that it involves making a practical judgement – a judgement about what it is best to do. And these judgements are typically *conscious* ones: the evidence for the existence of akrasia is precisely that we are sometimes aware of acting against our conscious judgements. An explanation of akrasia should thus include an account of what conscious practical judgements are and what cognitive role they have – a condition which not all existing theories meet. Here supermind theory can supply what is needed. The theory represents conscious practical

judgements as the conclusions of episodes of supermental practical reasoning in which we deliberately calculate what to do in the light of our premises and goals, applying the inferential techniques described earlier and prioritizing or revising our goals as necessary in order to resolve conflicts between them. And this gives us a new perspective on akrasia. For as I argued earlier, the judgements which issue from supermental reasoning influence our behaviour in virtue of our basic beliefs and desires about them. We act upon the conclusions of our supermental reasoning because we believe that these conclusions are dictated by our premising policies and want to adhere to those policies (as before, I use the term 'premising policies' to include both acceptances$_p$ and goal pursuits). That is, the efficacy of the supermind depends on the possession of a strong basic desire to adhere to one's premising policies. This desire can be overridden, however. In some cases the desire to adhere to one's premising policies will be outweighed by an even stronger desire to perform some other action. Such cases, I suggest, are precisely ones of akrasia.

An example may help to illustrate this. Suppose that I am debating what to do tonight. The late movie on television is one I have wanted to see for some time, but I am feeling tired and have to be up early for an important meeting. I review my goals, including seeing the movie, making a success of my job, and staying healthy. Prioritizing and making some simple inferences, I judge that the best thing is to get an early night. Since I have a strong basic desire to adhere to my goals and acceptances$_p$, I now become disposed to act upon this judgement. Yet as I am preparing for bed, I idly switch on the television and become engrossed in the movie. I do this, moreover, without revising my judgement. My basic desire to continue watching the movie outweighs my basic desire to act upon my judgement, but does not lead me to revise that judgement. I still judge it best to go to bed, and the awareness that I am failing to act upon this judgement causes me uneasiness and subsequent regret. (Note that in this example, performance of the akratic action was consciously considered and rejected prior to the formation of the judgement – seeing the movie was a goal of mine – but this is not an essential feature of the account. In other cases the akratic action may never have been consciously considered.)

I think that this view can resolve some of the puzzles surrounding akrasia and explain our competing intuitions about the condition. To begin with, it yields a straightforward explanation of why the agent performs the akratic action. Take the example just described. I stay up because I have a strong basic desire to see the movie – stronger than my desire to perform any of the alternative actions open to me. So the account vindicates the claim that akratic agents have reasons for their actions – that their actions are intentional. Yet the account also explains why we feel there is a sense in which such actions are performed without reason. For they are intentional *only at the basic level*. I have no supermental reason for staying up (although the goal of seeing the movie did figure in my conscious reasoning, it was overridden by other considerations). In fact, my action has no supermental explanation at all – the relative weakness of my basic desire to adhere to my premising policies having rendered my supermental states temporarily impotent. (Thus it is wrong to talk of a defeated desire

usurping behavioural control, as if the state which generates the akratic action is the same as that which was defeated in practical reasoning. The two states are quite different: the former a basic desire, the latter a superdesire.) It is important to emphasize that I am not claiming that any action which lacks a supermental explanation counts as akratic. Many ordinary nonakratic actions are generated without supermental activity and have only basic-level explanations. What renders an action akratic is not simply the *absence* of supermental influence, but the *failure* of it. Actions are akratic if they are performed in the face of a countervailing supermental judgement. Ones performed in the *absence* of such judgements are unreflective, but not akratic.

This view also allows us to reconcile the possibility of akrasia with Davidson's two principles, which now turn out to be ambiguous. Take the first principle – that if an agent judges that it would be better to perform action A rather than action B, then they want to perform A more than B. If we identify judgements with the conclusions of episodes of supermental practical reasoning, as I have suggested, then the truth or falsity of this principle depends on whether the phrase 'want more' refers to a basic preference or a supermental one. If the former, then the principle does not come out true. As we have seen, I can judge (consciously, at the supermental level) that it is better to perform A rather than B, yet have a stronger *basic* desire to perform B. If, on the other hand, the principle refers to a supermental desire, then there is a sense in which it does come out true. Of course, since superdesires are flat-out states, they cannot differ in strength in the way that basic desires can: one either pursues a certain goal or does not. However, goals can be assigned relative priority, and in this sense one can superdesire one thing more than another. And on this reading the principle comes out true, at least for the most part. If I judge that it is better to perform A rather than B, then I shall typically assign the goal of A-ing priority over that of B-ing. Davidson's other principle is also ambiguous. Davidson claims that if an agent wants to perform action A more than action B, and believes that they are free to perform either, then they will intentionally perform A if they intentionally perform either. Now, if the desire here is a basic one, then the principle is true: other things being equal, the stronger basic desire prevails. But there is no incompatibility with akrasia, since on the corresponding reading the other principle came out as false. If the desire is a superdesire, on the other hand, then the second principle is not true. I can assign the goal of A-ing priority over the goal of B-ing, yet perform B all the same. This will happen if my basic desire to perform B is stronger than my basic desire to act on my premises and goals – that is, if I am akratic. Again, then, there is no incompatibility with akrasia. It is true that if I superdesire to perform A more than B, then *in so far as my subsequent action is guided by my supermental states*, I shall perform A if I perform either. That is to say, understood as a claim about action that is intentional at the supermental level, the principle is true. There is, then, a non-equivocal reading on which both principles come out as true. There is still no incompatibility with akrasia, however. For all that is ruled out on this reading is the possibility of action that is both

akratic *and* intentional at the supermental level. And I have argued that akratic action is not intentional in this way.

The proposed account also highlights the nature of the irrationality involved in akrasia. It is simply a failure to do what one's conscious reasoning tells one to do. (If we think of conscious practical judgement as constituting the *will*, then the other name for akrasia – *weakness of will* – is an appropriate one for this failure.) However, the account does not require us to attribute any *basic-level* irrationality to the akratic agent. Since their basic desire to perform the akratic action is stronger than their basic desire to act upon their premising policies, it is rational for them to prefer it. The account is thus compatible with an austere view of the basic mind, which treats rationality as a constitutive principle of psychological interpretation. This is not to say that the akratic agent's basic attitudes are immune from criticism, however. For it may be unwise to value short-term pleasure above adherence to one's premising policies. There is little point committing oneself to a policy unless one is reasonably confident that one will adhere to it, and a persistent failure to adhere to past policies will erode this confidence. People who are routinely akratic may thus become unable to sustain an effective and coherent supermind (or, more precisely, unable to sustain a supermind capable of effective *practical* reasoning; they might still be able to maintain its theoretical functions).

To sum up, then, supermind theory can provide a robust account of akrasia, compatible with a commitment to basic-level austerity and without the extravagance of a partitioning strategy. Of course, the account applies only to agents with superminds, and it may be objected that this makes it implausible. Surely, we can imagine akrasia occurring in a single-level mind? This is to misconstrue the account, however. It is not offered as a conceptual analysis of akrasia, but as an empirical theory of *human* akrasia. I am not claiming that akrasia must involve supermental processes, merely that it does in us. The objector may say that even this is too strong. Could we not be akratic simply in relation to our non-conscious basic-level decisions, without any involvement of the conscious supermind? I grant that this sort of akrasia is conceivable (at least if we are prepared to give up an austere view of the basic mind), but see no evidence for its existence. As I mentioned earlier, the evidence for the existence of akrasia is simply that we are sometimes aware of acting against our conscious judgements.[1]

[1] Cohen also appeals to the belief/acceptance distinction in order to explain akrasia. However, his account is the opposite of the one outlined here – representing the akratic action as acceptance-based and the ineffective judgement as belief-based. According to Cohen, the akratic agent 'has a moral belief that requires him to bring it about that not-*p*, while he self-indulgently accepts as his maxim the principle of bringing it about that *p*' (1992, p. 153). This view seems to be dictated by Cohen's assumption that we are morally responsible for an action only if it stems from an act of voluntary acceptance. However, it seriously mischaracterizes the phenomenon. Typically, we do not deliberately *accept* that we shall perform our akratic actions – let alone accept maxims which dictate their performance. Rather, we lapse into them despite ourselves.

1.2 Akrasia and intention

So far I have been concerned with the standard form of akrasia, which consists in failure to act upon a practical judgement. It has been suggested, however, that other kinds of akrasia are possible, and, in particular, that we may akratically fail to form, or to act upon, an *intention* (Mele 1987a, 1992; Rorty 1980). In this section I shall say a little about this kind of akrasia. This will also give me the opportunity to say something about intention itself and how it fits into the two-level framework I have been developing.

We use the term *intention* to characterize both actions and states of mind (Bratman 1987). We speak of actions being *done with* intentions and of agents *having* intentions to perform future actions. In what follows I shall be using the term in the latter sense. My concern is with future-directed intentions, since it is in connection with these that the possibility of akrasia arises, and I shall say nothing about intentions manifested in action. (Thus, I shall not address the question of whether acting *with* an intention must involve *having* an intention.) Henceforth, 'intention' should be understood to mean 'future-directed intention', unless otherwise indicated.

It is sometimes claimed that intentions are just complexes of beliefs and desires – that to have an intention to perform an action is simply to have certain beliefs and desires relating to the proposed action. In recent years, however, a powerful case has been mounted for regarding intentions as a distinct species of mental state, with a distinctive functional role (Bratman 1987; Harman 1986; Mele 1992). Of these accounts, Michael Bratman's has been particularly influential. According to Bratman, to have an intention to do something is not simply to desire to do it or to believe that one will do it, but to be *committed* to doing it. Intentions, he argues, are parts of larger action-plans and help us to organize our activities over time and to co-ordinate them with those of others. Intentions can exercise this role, Bratman claims, because they have three features: they are conduct-controlling (if I intend to A now, then I shall normally at least try to A now; in this respect intentions differ from desires, which are only potential influencers of action); they have stability (once formed, intentions tend to resist revision); and they dispose us to reason in ways that will secure their satisfaction (to think about means and preparatory steps, and to ensure their consistency with other intentions one has) (Bratman 1987, pp. 16, 108–9). Bratman refers to this as the *planning theory* of intention.

This account is, I think, an attractive one. And it is attractive to think of intentions, so conceived, as supermental states. The planning and coordinating of action typically occurs at a conscious level, and the idea that intentions involve commitment suggests that they are flat-out, personally controlled states, like other supermental ones.[2] I suspect, then, that intentions belong exclusively to the

[2] Indeed, the planning theory of intention and the account of the supermind developed here complement each other well. Bratman notes that the planning theory assumes that planning takes place against a background of flat-out beliefs, and that the theory therefore needs to be supplemented with an account of the nature of such beliefs and of their relation to degrees of confidence (Bratman 1987, pp. 36–7 and p. 166).

supermental level and have no basic-level counterpart. I shall not argue for this view, however, and shall remain officially agnostic about the existence of basic-level intentions. Since we are interested in intention-related akrasia, and since the evidence for this relates to conscious intentions, we can set aside the question of whether there are non-conscious intentions. In what follows, 'intention' means 'conscious intention'.[3]

I suggest, then, that intentions are supermental states. More precisely, I want to suggest that they are a specialized form of goal pursuit, distinguished from the ordinary kind by two features. First, an intention aims at the performance of an action, or series of actions, rather than the existence of an independent state of affairs. Its content may be very specific, incorporating details of when and where and how the action is to be performed, and may become more specific as our planning proceeds. It is because intentions are directed at the performance of actions that they are directly conduct-controlling; desires, on the other hand, dictate actions only in conjunction with instrumental beliefs about how their objects can be realized. Their content also gives intentions a distinctive role in reasoning. Whereas the problem posed by a desire is primarily an instrumental one, that posed by an intention is primarily a planning one. The former requires us to think about which actions to perform in order to bring about the desired state of affairs, the latter about how to arrange our other activities in order to facilitate the performance of the intended action. The second distinguishing feature of intentions is that they have greater stability than other goal pursuits. Ordinary goals are open to revision or rejection at any time. It may be unwise to make a habit of continually changing one's goals, but there is no reason to regard any particular goal as specially resistant to change. An intention, on the other hand, can perform its function of facilitating long-term planning only if it is resistant to casual revision or rejection. Thus, to form an intention to A, one must not merely take A-ing as a goal, but also commit oneself to maintaining this goal unless strong reasons for changing it present themselves. To sum up: intentions, I suggest, are *stable action-oriented goal pursuits*.[4]

It may be objected that this account involves a regress. I claim that intentions are constituted by premising policies. But, surely, adopting a *policy* involves forming an intention – an intention to perform the actions required by the policy (see Bratman 1987, pp. 87–91; Bratman calls such intentions *general intentions*, since they are open-ended, rather than being directed at a specific action). So if intentions are policies, then forming an intention will involve forming a second intention, which will in turn involve forming a third, and so on. Any account which has this consequence must be

---

[3] I am happy to concede that a primitive kind of intention could exist at the *sub-personal* level. This might consist simply in the memory of the conclusion of an episode of practical reasoning, stored for subsequent execution when the time is right. But I suspect that full-blown intentions of the sort Bratman discusses are found only at the supermental level. And it is to such states, I think, that everyday talk of intentions refers.

[4] Let me stress that this is not offered as a conceptual analysis of intention, but simply as an hypothesis about how human intentions are constituted.

wrong. There are two lines of response to this. The first concedes that policies involve general intentions, but points out that the proposed account applies only to *conscious* intentions. I can maintain that forming a conscious intention involves forming a further general intention to pursue a premising policy, but insist that this general intention is a non-conscious, basic one, and that such intentions are constituted differently from conscious ones. This would, of course, involve conceding the existence of basic-level intentions. The second option, and the one I tentatively endorse, is to deny that policy adoption must involve forming a general intention. As I argued in chapter 4, to have a policy of A-ing it is sufficient to believe that one is committed to a such a policy, to know what the policy requires, and to desire to adhere to it. There is no need to invoke intentions in an account of policy adoption. This is not to deny that *some* policies may involve intentions. In particular, *conscious* ones typically will. (The policies that support supermental states, by contrast, are usually non-conscious.) To form a conscious policy of A-ing is, I suggest, typically just to form the conscious intention to pursue a policy of A-ing. On the proposed account, this intention will itself involve adopting a *non-conscious* policy of taking the execution of a policy of A-ing as a stable action-oriented goal. That is, conscious policies will be realized in non-conscious second-order policies.

Return now to akrasia. I suggested that it is possible to be akratic in the formation or execution of intentions. One might judge it best to perform some future action, yet akratically fail to form an intention to perform it. Or one might form the intention, yet fail to execute it properly – either omitting to do the necessary planning or simply failing to perform the action when the time comes. Now, if intentions are supermental states, then these kinds of akrasia are easily explained. The explanation is the same as for the standard kind: the agent's basic desire to perform some other action, or to do nothing, is stronger than their desire to adhere to their premising policies. For example, after I have decided that it is best to visit the dentist next week, my reluctance to make the necessary arrangements might override my basic desire to adhere to the policies that dictated the decision, leading me to refrain from adopting the making of the visit as a stable action-oriented goal. Or, after I have adopted that goal and made the initial arrangements, my fear of the dentist's chair might outweigh my desire to stick to my goals, with the result that I do not turn up for the appointment. As with the standard kind, this sort of akrasia need not involve any basic-level irrationality – the agent does what they desire most to do – though the preferences it manifests are open to criticism. In preferring immediate satisfaction over adherence to their policies, akratic agents will undermine the effectiveness of their supermental processes, making it harder for them to rely on those processes in the future.

Episodes of intention-related akrasia of the sort just described should be distinguished from cases in which an agent *changes* or *revises* their intentions without sufficient reason for doing so. Such cases are not ones of akrasia, properly speaking, since they do not involve acting *against* one's intentions, but they do display a sort of weakness of will. (Indeed, Holton argues that unreasonable revisions of intention are

the paradigm cases of weakness of will: see Holton 1999.) As I emphasized, it is important to persist in one's intentions, and a failure to do so will undermine their effectiveness. Incontinent intention revision might be thought of as involving a failure to adhere to a meta-policy of persisting in one's intentions.

2. Self-deception

We often talk of people *deceiving themselves* – that a partner is faithful, that they do not have a drink problem, that their failure at work is due to the envy of colleagues. Yet the very idea of self-deceit can seem paradoxical. If I deceive you, then I intentionally induce you to believe a proposition which I believe to be false. If self-deceit follows the same pattern, then a self-deceiver is one who intentionally induces themselves to believe a proposition which they believe to be false. But how can this happen? How can I believe that p while also believing that not-p? And how can I intentionally induce myself to believe a proposition I think is false? Surely, any attempt I make will be self-defeating – serving simply to draw my attention to the fact that I think that it is false?

There are various ways of responding to these problems. Some writers weaken the interpersonal model, arguing that self-deceivers do not really know the truth about the matter on which they are deceived. Self-deception, they claim, involves believing a proposition in the face of strong evidence for its falsity, but without actually believing that it is false (Canfield and Gustavson 1962; Mele 1983; Penelhum 1964). Others suggest the opposite: that self-deceivers do know the truth, at least non-consciously, but do not really come to believe the falsehood. To deceive oneself that p, they suggest, one need not actually believe that p, but simply be disposed to avow it sincerely (Audi 1982b), or to avoid entertaining the occurrent thought that not-p (Bach 1981). A third group of writers adopt a partitioning strategy, retaining the interpersonal model, but positing distinct subagents within the self-deceiver, one of which deceives the other (Davidson 1982; Pears 1984, ch. 5).[5]

None of these strategies is without its drawbacks. Accounts that weaken the interpersonal model, while they may describe real psychological conditions, arguably fail to come to grips with full-blown self-deceit. Partitioning strategies, on the other hand, have an ad hoc air about them and, if taken literally, involve positing subsystems with implausibly sophisticated motives, plans, and self-monitoring abilities (Johnston 1995). Again, supermind theory offers an alternative approach, which has the robustness of a partitioning strategy without its extravagance.

There are two ways in which the theory is particularly well placed to explain self-deception. It can account for the conflicting attitudes involved by assigning them to different levels, and it can explain the intentional element in the process by supposing

[5] This brief list is not exhaustive, of course, though is it representative of some of the main lines of response to the problem. For a survey of work on self-deception, see Mele 1987b, and for important collections of papers, see Martin 1985 and McLaughlin and Rorty 1988.

that the deceptively induced state is a supermental one, formed in response to basic-level desires. Here is what I suggest happens in a typical case. The agent has strong evidence for the falsity of some proposition, p, and a strong basic belief that not-p. However, they also have a strong basic desire that p, feel anxiety whenever they entertain the conscious thought that not-p, and are deeply unwilling to $accept_p$ that not-p and to face up to its epistemic and practical consequences. These desires and anxieties lead them to pursue what I shall call a *shielding strategy*, which involves manipulating their supermental processes in ways designed to keep the distressing thought at bay. In the weakest case they simply take steps to avoid the conscious thought that not-p (see Bach 1981 for a description of various strategies we can use for doing this). But in full-blown self-deceit they go a step further and adopt p as a general premise (that is, as a premise for deliberation on all relevant topics), thereby ending deliberation on the matter and committing themselves to a view they find comforting.

I claim, then, that full-blown self-deception involves a form of pragmatic general $acceptance_p$, in which a proposition is $accepted_p$ for non-epistemic reasons and regardless of the evidence for its truth. To this extent, the self-deceiver is like the positive-thinker who $accepts_p$ that they are confident and capable in order to boost their self-esteem. Where, then, does the element of deceit enter? It enters, I suggest, in the self-deceiver's attitude towards their $acceptance_p$. The positive-thinker may be fully aware that their $acceptance_p$ is pragmatically motivated and unsupported by evidence. The self-deceiver, on the other hand, does not consciously acknowledge this. They either do not explicitly consider the matter or, if they do, think of their attitude as one of justified belief. This is crucial, of course: to acknowledge that they had $accepted_p$ p without evidence would be to reopen the issue of whether p is really true, which is precisely what they want to avoid. Moreover, this lack of conscious awareness is not accidental, but motivated – sustained by a basic desire to maintain the shielding strategy. It is here that talk of deceit becomes particularly apposite. The self-deceiver unconsciously 'fiddl[es] with the evidential books', as Cohen puts it (1992, p. 145), focusing on evidence that favours p and ignoring or explaining away evidence that counts against it, thus sustaining the illusion that their $acceptance_p$ of p is epistemically motivated. Such activities will be particularly important at the time the $acceptance_p$ is formed – allowing the commitment to be made without obvious epistemic impropriety.

I suspect that the self-deceiver's basic desire to maintain their shielding strategy will have a further effect, too, contributing to a shift in their deliberative standards which makes their $acceptance_p$ appear more like a genuine belief. As I argued in chapter 5, $acceptance_p$ without high confidence is not really belief. An agent believes that p (in the strand 2 sense) only if they are disposed to take it as a premise in deliberations where they want to take only truths as premises ('TCP deliberations'). Now, by this criterion, the self-deceiver does not count as believing that p. Since their confidence in p is low, they would not take it as a premise in deliberations where they genuinely wanted to take only truths as premises – either suspending judgement or taking not-p as a premise instead. This would, of course, expose and undermine their

shielding strategy – forcing them to acknowledge that they had accepted$_p$ p for pragmatic reasons and without evidence. (Why else should they be so reluctant to rely on it when it is important to rely on the truth?) The self-deceiver's basic desire to maintain their shielding strategy is thus in tension with any basic desire they may have to treat a deliberation as TCP, and, if strong enough, may override the latter, leading them to reclassify the deliberation as non-TCP. For example, take a case where one is asked whether it is the case that p and has no reason for deceiving the questioner or concealing one's opinion. In normal circumstances the ensuing deliberation would count as TCP – one would want to take only truths as premises in deciding what to say. But if one is self-deceived with respect to p, then the situation may be different. The desire to maintain one's shielding strategy may override the desire to take truths as premises, leading one to continue premising that p, and so to declare that p. It is important to stress that the desires mentioned will be basic-level ones and the process will not involve any conscious deceit. At a conscious level the agent will simply entertain the thought that p in, as it were, *premising mode* – as something they are committed to taking as a premise – unaware that they are doing so in response to a non-conscious desire to maintain a shielding strategy, rather than out of a concern with truth. The self-deceiver will also typically avow belief in p and tell themselves that they believe it – again, promoted by a non-conscious desire to maintain their shielding strategy and without any conscious insincerity.

The upshot of this is that if the self-deceiver's basic desire to maintain their shielding strategy is strong – as it will be in cases where they find the thought that not-p very troubling – then there will be few, if any, deliberations which they regard as TCP (the exceptions being ones where what is at stake is something more important to them than p – life itself, perhaps, or the life of a loved one). In such cases, the self-deceiver will take p as a default premise in most of their deliberations, just as if they genuinely believed it, and their attitude to p will be hard to distinguish from genuine superbelief.

To sum up, then, I suggest that in a typical case of self-deceit, the agent (1) has a strong basic belief that not-p, (2) has a strong basic desire to avoid consciously acknowledging that p, and consequently (3) pursues a shielding strategy, which involves adopting p as a general premise, manipulating the evidence so as to prevent conscious acknowledgement that they do not really believe that p, and treating as non-TCP many deliberations which they would otherwise have treated as TCP.

This view offers solutions to the core problems associated with self-deceit. Since the opposing attitudes involved are located at different levels, they do not come into direct conflict with each other, and since the shielding strategy is motivated at a non-conscious level, it is not self-defeating, as it would be if it were conscious. (Indeed, preventing conscious awareness of its own existence is an essential part of the strategy.) Moreover, as with akrasia, the account does not require us to attribute any basic-level irrationality to the self-deceiver and is thus compatible with an austere view of the basic mind. Accepting$_p$ p does not involve believing p at the basic level, and though it does involve possessing certain basic beliefs – among them, that one has

accepted$_p$ p – none of these is inconsistent with the self-deceiver's strong basic belief that not-p. The other aspects of a shielding strategy are also compatible with basic-level rationality. The self-deceiver manipulates the evidence and adjusts their deliberative standards, but the effects of these activities are confined to the supermental level, and it is quite rational for the self-deceiver to engage in them, given their strong basic desire to avoid consciously acknowledging the truth. (As with akrasia, this is not to say that the self-deceiver's basic attitudes are beyond criticism. It may be unwise to place a high value on shielding oneself from unpleasant truths, and better in the long run to face up to them.)

I am not going to defend this account here, but I shall briefly address a couple of objections to it before moving on. First, it may be objected that the account does not capture full-blown deceit, since it represents the self-deceiver, not as *believing* that p, but only as *accepting*$_p$ it, and thus weakens the interpersonal model from which we started. My response here is to point out that the weakening involved is slight.[6] For even if the self-deceiver's attitude to p is not belief, it is very *like* belief and quite different from the sort of general acceptance$_p$ whose status is openly acknowledged. As we saw, a self-deceiver will regard themselves as believing that p if they consciously consider the matter, will avow p without any conscious insincerity, and will take p as a default premise in all or most of their relevant conscious deliberations, including many which they would otherwise have regarded as TCP. Indeed, it may be that their attitude does in fact fall within the extension of the folk concept of belief and thus constitutes an exception to the claim that high confidence is necessary for belief. (This would be in the spirit of our original definition of superbelief as unrestricted acceptance$_p$. Since the self-deceiver is reluctant to treat deliberations as TCP, their acceptance$_p$ of p will in practice be almost completely unrestricted.) At any rate, there is a real tension between the self-deceiver's conscious and non-conscious attitudes, and given the non-conscious manipulation involved in supporting the former, it seems quite appropriate to liken their condition to interpersonal deceit.

Secondly, it may be objected that not all self-deceitful thoughts are comforting ones which serve to shield us from distressing truths. We can also deceive ourselves into thinking painful thoughts, as when a jealous person deceives themselves into thinking that their partner is unfaithful, despite having no evidence that they are (Davidson 1985, p. 144). I grant that such cases exist. The proposed account can, however, easily be extended to deal with them. The general nature of the deceit is the same as in the standard case: the agent accepts$_p$ the distressing thought as a general premise and then non-consciously manipulates the evidence and adjusts their deliberative standards in order to prevent themselves from consciously acknowledging what they have done. The difference is primarily one of motivation: the aim is not to shield oneself from a distressing thought, but to *expose* oneself to one – whether for masochistic reasons or in order to pre-empt future distress.

---

[6] All accounts involve *some* weakening of the interpersonal model. In interpersonal deceit the deceived person believes that p while the deceiver does not, and it is impossible for a self-deceiver simultaneously to believe that p and not believe that p – at least in the same sense.

Finally, a note on the relation between self-deceit and wishful thinking. The latter, I suggest, also involves accepting$_p$ a proposition for pragmatic reasons and then non-consciously manipulating things in order to cover one's tracks. What distinguishes it from self-deceit, I suggest, is the degree of confidence the agent has in the accepted$_p$ proposition and the extent of the subsequent manipulation required. In self-deceit the agent's confidence in the accepted$_p$ proposition is very low, and the need for subsequent manipulation correspondingly high. In wishful thinking, on the other hand, the agent's confidence is somewhat higher, though still not high enough for normal belief, and the need for manipulation less.

3. First-person authority

Another area of application for supermind theory lies in the explanation of first-person authority, where the theory supports a *performative* account of the phenomenon. In this section I shall briefly outline this account and address some objections to it.

3.1 First-person authority as performative

A person's claims about their own current mental states ('avowals') are usually regarded as authoritative. This is not to say that we would never question them – we might suspect that the speaker is lying or that they are guilty of self-deception. But we do assume that people cannot make straightforward errors about their own mental states, through overlooking or misinterpreting the evidence, as they might in describing their non-mental states. Indeed, avowals do not seem to be made on the basis of observation or interpretation at all; in typical cases we can say what mental states we have straight off, without first checking the evidence. (Or, at any rate, this is a very common assumption. There is, in fact, evidence that people sometimes *confabulate* when asked to report their beliefs and desires, yet without realizing that they are doing so. I shall say something about these cases later.)

First-person authority is notoriously difficult to explain – at least if we reject the idea that each person's mind is a distinct non-physical realm to which they have unique and infallible access. If our mental states belong to the public world of physical causes and effects, then how can we be specially authoritative about them – any more than we can about any other aspect of the physical world?[7] It is sometimes suggested that we possess an *inner sense* – a self-scanning mechanism which gives us specially reliable access to our mental states (Armstrong 1968; Goldman 1993). However, first-person authority seems to involve something more than reliable access, and it is often suggested that it has a conceptual character. There are various ways of fleshing out this idea. According to *expressivist* theories, avowals serve to express mental states, rather

[7] The problem becomes even more difficult if we hold that the content of mental states is determined in part externally, by their causal relations to features of the external world. For then it seems that authority about our mental states will require a corresponding authority about aspects of the external world, too. For present purposes I shall set aside this aspect of the problem.

than to report them, and thus cannot misreport them (Wittgenstein 1953, 1980). *Constitutive* theories, on the other hand, allow that avowals are reports, but deny that they are empirical ones. According to such theories, in the right circumstances, believing oneself to possess a certain mental state makes it the case that one does possess it: there is nothing more to possessing the state than conceiving oneself to do so. Thus, in the right circumstances, sincere avowals are true a priori (Wright 1989, 1998). Functionalists can make a related move, claiming that it is part of the functional role of mental states to generate accurate second-order judgements about themselves, and part of the functional role of such judgements to be caused only by the states they are about – thus making it a priori that such judgements are true (see Fricker 1998). These views have their attractions, but also some well-known weaknesses. Expressivist theories have the consequence that first-person uses of mental-state terms have a radically different meaning from other uses; constitutive theories tend to assume an irrealist conception of mental states; while the functionalist theories mentioned impose very stringent conditions for belief possession. Again, supermind theory offers an alternative approach, which also represents first-person authority as having a broadly conceptual character and which has a number of further attractions, too. Note, however, that the suggestion will apply only to beliefs, desires, and intentions. Perception and sensation will require a different treatment.

An important feature of first-person authority is that it holds only for *conscious* mental states. In order to identify my non-conscious mental states I shall need to observe and interpret my behaviour, and the conclusions I arrive at will not be authoritative. My close friends may be better observers of my behaviour than I am. (Thus avowals are authoritative only if intended as ascriptions of conscious mental states; henceforth I shall take this limitation as read.) Now, I have argued that conscious beliefs and desires are supermental states, and I want to suggest that the authority we have in regard to them depends on distinctive features of the supermind. The key feature is that supermental states are under personal control. We can actively adopt superbeliefs and superdesires by committing ourselves to appropriate premising policies. And I suggest that avowals serve to make or reaffirm premising commitments of this kind. A sincere utterance of 'I believe that p', used in reference to a conscious belief, makes a commitment to accepting$_p$ p unrestrictedly (that is, to taking it as a default premise in all relevant deliberations, including TCP ones). It is, in effect, shorthand for 'I hereby accept$_p$ p unrestrictedly.' Similarly, an utterance of 'I want x', used to refer to a conscious desire, makes a commitment to taking x as a goal. In these cases, the primary function of the avowal is neither descriptive nor expressive, but *performative*.

This approach has a number of attractive features. In order to explain I shall need to say a little about how I view performatives. (It is not essential that the reader share this view; the assimilation of avowals to performatives would, I think, remain attractive on other views of performatives, though its attractions stand out particularly well on this one.) A performative utterance is one which performs the action it apparently describes – for example 'I thank you' or 'I promise to be there.' It is

sometimes denied that such utterances have truth values, but I shall assume otherwise. A performative utterance, I shall assume, not only performs some act, but also states that the speaker performs it. This case for this view is strong (see Bach and Harnish 1979; Heal 1974; Searle 1989). I shall also assume that performative utterances derive their status from the intentions with which they are made, rather than from special meanings attaching to the terms used, or from special conventions surrounding their use (though some formal performatives are dependent on social conventions). Again, there is a strong case for this view (see Bach and Harnish 1979, 1992). (Note that I am here using the term 'intention' in its other sense, to characterize actions rather than states of mind. There is no implication that performative utterances must issue from previously formed *future-directed* intentions.)

Now, it is a consequence of the view of performatives just outlined that performative utterances are self-guaranteeing. In sincerely uttering the sentence 'I promise to A', I make a promise to A, and thus bring it about that my utterance is true. And if avowals are performatives, then they, too, will be self-guaranteeing. In sincerely uttering 'I believe that p', I accept$_p$ p unrestrictedly, and thus bring it about that I believe that p; the utterance simultaneously asserts that I am a p-believer and makes me one. The performative view thus explains the authority of avowals as a species of the more general phenomenon of self-guaranteeing performativity. This account of first-person authority might loosely be called a *constitutive* one, in that it holds that a sincere avowal makes it the case that the speaker possesses the avowed state; but it should be clearly distinguished from the constitutive theories mentioned earlier. According to those theories, simply *believing* oneself to possess a certain mental state is, in the right circumstances, constitutive of the state. On the view outlined here, by contrast, it is not the *belief* that one possesses a mental state that is constitutive of the state, but the *act* of committing oneself to a suitable premising policy. The mental state is constituted by the policy and the avowal is constitutive of the mental state only because it makes a commitment to the policy.

The performative view explains other features of avowals, too. Consider, first, a feature of their epistemology. It is often noted that if asked whether we believe a certain proposition, we direct our attention, not to ourselves and our mental states, but outward, to the aspects of the world that make the proposition true or false (Carruthers 1996c; Evans 1982, p. 225; Gordon 1995). If asked whether we believe that beef is safe to eat, we think about beef, not about ourselves. (This does not contradict the earlier claim that first-person belief reports are not based on evidence, since the evidence in question is quite different. The point was that such reports are not based on evidence for the existence of the *beliefs* reported.) Similarly, if asked whether we want a certain outcome or intend a certain action, we think about the desirability of the outcome or the value of the action, not about ourselves. Now, if avowals are performatives, then it is easy to see why this is so. For in avowing belief in a proposition, we are accepting$_p$ it unrestrictedly, and since it will be rational to do this only if we think the proposition likely to be true, it is appropriate to consider the evidence for it before committing ourselves. Similarly, in saying that we desire a

certain outcome or intend a certain action, we are committing ourselves to pursuing the one or performing the other, and it is appropriate to consider their utility before doing so.

The view also resolves a puzzle about the semantics of psychological terms. On the one hand, first-person present-tense uses of psychological verbs seem to possess some special semantic feature which accounts for their authority. On the other hand, we have a strong intuition that psychological verbs have the same meaning in all their various persons and tenses. If I say that I currently believe a certain proposition, then I seem to be saying exactly the same of myself as I would be saying of you if I said that you believed it, or of my past self if I said that I formerly believed it. Now, if avowals are performatives, then this tension is explained. For performative verbs have the same meaning in all their uses, but can be used performatively only in the first-person present tense. My utterance of 'I promise to A' says the same of me as my utterance of 'You promise to A' says of you (namely, that we promise to A), but only the former utterance additionally *makes a promise* to A. That is to say, first-person present-tense uses of performative verbs are special, not because they have a different meaning, but because they perform an additional action. The same, I suggest, goes for the psychological verbs 'believe', 'desire', and 'intend'. When I say that I believe that p, I am saying the same of myself as I say of you when I say that you believe that p, but when I say it of myself I do something extra – namely, commit myself to a premising policy – which makes the statement true.

3.2 Objections and replies

This is not the place for a full defence of this account, but I shall briefly consider some possible objections to it, in order to help fill in the picture.

First, it may be objected that on this account all we can be authoritative about is that we are currently committing ourselves to a premising policy; we cannot be sure that the commitment will produce a continuing allegiance to the policy and thus a persisting superbelief or superdesire. We might make the commitment and then immediately forget that we have done so or suddenly lose the will to discharge it. This is true, but does not amount to a serious limitation on first-person authority. For in normal circumstances a commitment *will* generate a continuing allegiance of at least some duration. Instant amnesia is rare, and whatever reasons we have for making a commitment will be reasons for discharging it too. Only in pathological cases, then, will a sincere avowal fail to introduce the corresponding mental state, and it is no surprise that first-person authority may fail in such cases. (This is not to say that we never forget or abandon our premising policies, just that we do not do so immediately upon adopting them. An avowal does not, of course, guarantee that the avowed state will never be lost.)

A second objection is that the proposed account applies, at best, only to *some* avowals. Some avowals may make premising commitments, but, surely, many others serve simply to report commitments already made – that is, to let others know what we believe or want or intend. And in these cases first-person authority will not hold,

since we may misremember our commitments. If asked whether I believe a particular philosophical theory which I thought about at some point in the past, I may misremember the conclusion I came to – perhaps thinking that I decided to $accept_p$ it when in fact I did not. Now, it is certainly true that not all avowals serve to make *new* premising commitments; many, I agree, reflect commitments already made. But even in these cases, I suggest, the avowal still involves commitment. In avowing a mental state we report that we *currently* possess it – that is, that we are currently committed to the relevant premising policy. And we are not bound to make such a report, even if we are aware of possessing the state in question up to this very moment. For we might change our minds right now – that is, revise or repudiate the premising policy involved. Indeed, being questioned about our beliefs or desires may itself provoke such a change ('Well, I did believe it, but now you ask, I'm not so sure'). The decision to avow a mental state, then, reflects a tacit decision *not* to change our minds, and *recommits* us to the relevant policy. Thus, in the imagined case, if I were to avow belief in the philosophical theory under the misapprehension that I had previously $accepted_p$ it, then I would *thereby* $accept_p$ it and incur the relevant premising commitments.

Thirdly, it may be objected that for a speech act to constitute a commitment to a premising policy it has to be intended as such, and we simply do not think of avowals in this way. This objection is weak, however. For while it is true that we do not consciously think of avowals as commitments to premising policies, it is possible that we do so *non-consciously*, at the basic level. This is not to say that avowals are never made with conscious thought, just that their function is dependent on more specific, non-conscious attitudes. At a conscious level, we simply intend to say what we believe; but saying what we believe, I claim, involves producing an utterance with the non-conscious intention of making a commitment to a premising policy. This intention reveals itself in our subsequent behaviour – in the way that we feel bound to treat the content of the avowed belief as a premise.

This objection prompts a fourth. If the function of avowals depends on our non-conscious attitudes, does this not compromise our title to first-person authority? We might think that we were making a sincere avowal when in fact our utterance was motivated by a non-conscious desire to deceive our hearer, rather than by a wish to make a serious premising commitment. This objection is also weak. For in so far as it trades on the possibility of non-conscious deceit, it applies to *any* theory of first-person authority. On any account, first-person authority holds only for avowals that are sincere, in the sense of not deceitful, and if we allow that non-conscious deceit is possible, then we must allow that some avowals may be insincere, and hence non-authoritative, even though no conscious deception is intended. If that conclusion is rejected, then the problem lies with the claim that non-conscious deceit is possible, rather than with the account of first-person authority offered here. (In practice, non-conscious deceit of this kind would soon reveal itself to the deceiver; we would soon realize that the avowal we had made was not sincere when we found ourselves unwilling to take the content of the avowed attitude as a premise in our private reasoning.)

Similar considerations apply to self-deceit. I claim that sincerely to avow belief in a proposition is to $accept_p$ it unrestrictedly – that is, to commit oneself to taking it as a default premise in all relevant deliberations, including TCP ones. Sincere avowals will thus require high confidence, since this is a prerequisite for premising in TCP deliberations. And, again, we cannot be sure that any given avowal we make is sincere in this way. For all we know, it might have been motivated by a non-conscious desire to maintain a shielding strategy, and the commitment involved might have been restricted to non-TCP deliberations and unaccompanied by high confidence. We may, in other words, be deceiving ourselves. Again, however, this does not constitute a special objection to the present proposal, since on any account, first-person authority will fail when we are self-deceiving. (I am assuming here that self-deceivers do not count as genuinely believing the propositions about which they are deceived. If that assumption is false – and, as we saw earlier, it may be – then the objection does not arise.)

The moral of these considerations is that the authority of avowals is defeasible and, more specifically, that the defeating conditions can be non-conscious. It follows that we cannot be sure that any given avowal we make is authoritative. That is to say, we are not authoritative about the existence of the conditions under which we are authoritative about our mental states: we do not have second-order authority of that kind. But it remains true that we possess first-order authority: in the right circumstances – circumstances which only rarely fail to obtain – we can pronounce directly and authoritatively on our current conscious mental states, without risk of error through mistaking or misinterpreting the evidence.

Finally, I want to consider a challenge to the very existence of first-person authority. There is evidence that our verbal reports of our mental states and processes are sometimes confabulated. Psychologists have found that it is possible to influence a person's behaviour by means of suggestions and other stimuli which are not obviously relevant to it. When the subjects of these experiments are asked to explain their behaviour, their responses are surprising. They do not mention the stimuli or confess ignorance of their motives, but instead invent some plausible reason for their actions (Nisbett and Wilson 1977).[8] What is happening in these cases, it seems, is that the subjects' actions are guided by non-conscious processes of which they are unaware, and their reports are attempts to interpret their own behaviour as if from a third-person perspective. So far, there is no conflict here with first-person authority; as I emphasized earlier, we are not authoritative about our non-conscious mental states. The problem is that in these cases the subjects do not realize that their reports are

[8] Similar results occur with subjects who have undergone commisurotomy (surgical severing of the corpus callosum which connects the two halves of the brain). If a command such as 'Walk!' is presented to such a subject in the left half of their visual field, so that it is received only by their right hemisphere, then they will typically comply with it. Yet when asked to explain their action, they will not mention the command, since their left hemisphere, which controls speech, has no information about it. Instead, they will invent some plausible reason for the action, such as that they were going to get a drink (Gazzaniga 1983, 1994).

merely self-interpretations – they make them with complete sincerity, just as if they were reporting conscious mental states and processes. So, it seems, avowals can be erroneous. Indeed, for all we know, many of our avowals may be like this, and first-person authority an illusion.

I think that the threat here is specious. The main point to stress is that first-person authority in the strict sense holds only for one's *current* mental states, and the reports in question are of past ones – of the states and processes that led to a prior action. Now, it is true that the actions are not *long* past, and the subjects seem to indicate that they still possess the beliefs or desires they cite. But it is not clear that they are wrong about this. For in avowing a belief or desire, I suspect, they will silently endorse it, committing themselves to taking it as a premise and to regulating their future behaviour accordingly. So although they did not previously possess the mental states they mention and are wrong to cite them in explanation of their earlier actions, they do *now* possess them, at the supermental level, and their avowals are therefore accurate. These cases thus present no threat to the existence of first-person authority, at least as understood within the context of supermind theory. (What these cases do show is that we know less than we think about our mental processes. But this is, I think, just because we overestimate what we *can* know about them – perhaps because we tacitly subscribe to the unity of belief assumption. We are, I assume, usually able to give accurate reports of the reasons for those of our actions that are motivated at a conscious level. Our mistake is to think that we can give similarly accurate accounts of all our actions and to assume that any explanations that come to mind must be true.)

This is all I shall say here in defence of the proposed account. The general idea that avowals are performatives has been canvassed before – most recently by Jane Heal (Heal 1994, 2002). Heal arrives at the view by abductive means, arguing first for a broadly constitutive account of first-person authority, and then developing the idea through an extended comparison between avowals and promises. The picture that emerges is similar in outline to the one proposed here: in self-ascribing the belief that p we also judge that p is true, thereby introducing the ascribed belief. Heal does not flesh this account out in any detail, but it is interesting to note that she gestures at the need for a two-strand theory of belief – drawing a distinction between what she calls *natural beliefs* and *personal beliefs*, the former being non-conscious states over which we have no special authority, and the latter, states of a more reflective kind for which a performative account is appropriate. This suggests that Heal's view may harmonize well with the one developed here.[9]

[9] There is another anticipation of the present account in the 'constructivist' theory of self-knowledge proposed by Julia Tanney (Tanney 1996). Tanney suggests that psychological self-description is to some extent a creative enterprise: within certain limits, we are to free to choose which self-interpretations to accept – our choices carrying a commitment to act in accordance with the interpretations chosen.

4 Scientific psychology

So far, I have focused on applications within the scope of folk psychology: the phenomena considered have been ones characterized in folk-psychological terms and recognized in everyday discourse about the mind. The ease with which supermind theory can explain these phenomena is further evidence that the theory is implicit in folk psychology. But if the theory is, in addition, *true* – as I have claimed it is – then it should also have applications outside the realm of folk psychology. The theory should illuminate, and be illuminated by, work in various areas of scientific psychology. Exploring these connections will be a separate task, but I shall briefly mention some links here and suggest how they might be developed.

4.1 Dual-process theories

The most obvious link is with 'dual-process' theories of reasoning developed by some psychologists working on reasoning and rationality. Experimental evidence shows that humans consistently make certain basic errors in reasoning tasks, particularly ones involving conditionals and probabilities. These errors seem to reflect certain innate cognitive biases, and subjects continue to make them, particularly when a fast response is required, even when they have been taught the principles governing correct responses and shown themselves able to apply them in other situations. Several writers have suggested that this indicates that humans possess two reasoning systems: a non-conscious system ('system 1'), which is fast, inflexible, and automatic, and which relies on heuristics rather than rules of inference, and a conscious system ('system 2'), which is slow, flexible, and controlled, and which is responsive to verbal instruction (Evans and Over 1996; Sloman 1996; Stanovich 1999). It has been suggested that individual differences in cognitive ability are largely due to system 2 processes, while system 1 processes display little individual variation (Stanovich 1999).

There is a clear correspondence here with the two-level theory set out in this book, with system 2 corresponding to the supermind and system 1 to the basic mind, and I believe that supermind theory can be fruitfully integrated with dual-process theories. In particular, the account of the supermind developed here may provide a framework for thinking about system 2 processes – explaining how they are constituted, how they influence action, and the nature of the states upon which they operate. Experimental work may in turn suggest ways in which the account can be revised and refined. The relation between the basic mind and system 1 is more complex, and at first sight there may seem to be an incompatibility here. The basic mind, as pictured here, is a collection of multi-track behavioural dispositions (thickly carved functional states), whose attribution carries a presupposition of rationality. System 1, on the other hand, is conceived as a processing system, or suite of processing systems, whose operations sometimes violate rational norms. But the conflict here may be only superficial. We can think of the psychologist's account of system 1 as a description of the sub-personal processes underlying the basic mind. That is to say, we can regard both accounts as focused on the same level of cognition – the folk account serving as a

competence theory, which provides an idealized description of the system's behaviour, and the psychologist's as a performance theory, which aims to identify the underlying mechanisms and their particular idiosyncrasies. (This is what we should expect: the basic mind is a folk construct, and the folk have neither great interest in non-conscious cognition nor access to the experimental data needed in order to theorize about it.)

4.2 Evolutionary and developmental psychology

Another broad area of application for supermind theory lies in thinking about the development of the human mind, both in the species and in the individual, and the extent to which our mental capacities are innate. There are obvious implications here for supermind theory and, in particular, for its picture of the conscious mind. If the conscious mind is a virtual structure, as I have suggested, formed by the exercise of various personal abilities, then in order to understand its development we shall need to think about how these abilities develop and how they become co-ordinated in the service of supermentality. The issues here are complex, but I shall offer a few suggestions.

First, I suspect that the elements of supermentality developed independently in the species and continue to develop independently in the individual, only later being co-opted to support a supermind. Consider some of the abilities required for simple language-based forms of acceptance$_p$ and goal pursuit. In addition to natural language, these include: (1) private speech – the habit of speaking to oneself; (2) meta-representational abilities – the ability to think of sentences as representations of states of affairs, actual and non-actual; (3) personal inferential skills, including the ability to construct explicit arguments following learned inference rules; (4) meta-cognitive skills – skills in focusing attention, searching memory, keeping track of information; and (5) strategic abilities – the ability to adopt and execute policies of action. It is also likely that supermentality requires theory of mind. Pursuing a premising policy involves having beliefs about propositions and the attitudes one has adopted towards them. And if theory of mind is required for such beliefs (as it surely is), then it will be required for supermentality. Now, each of these abilities has a more basic function than that of helping to support a supermind. Private speech can help to focus attention and aid memory; meta-representational abilities are required for reading, writing, and other sophisticated forms of language use; inferential skills are needed for rational debate with one's peers; meta-cognitive skills can improve performance on a wide range of tasks; strategic abilities are required for engagement in structured social activities where there is division of labour; and theory of mind is, of course, essential for efficient social interaction. These skills are ones that children acquire in the course of normal development, and there may be an innate component to them.

This suggests, then, that the conscious mind is a *kludge* – a jury-rigged system assembled from pre-existing components designed for other purposes. But how do we come to unite these disparate skills in the service of supermentality? Does each of us need to learn the trick of premising, or are we innately disposed to master it? I suspect

the latter, thanks in part to what is known as the *Baldwin effect*. The idea is that once a useful skill has been discovered – say, the knack of making a certain tool – then there will be pressure for other members of the discoverer's community to acquire it too. Those who find it easy to acquire the skill will thus have a selective advantage over those who find it hard, and individuals who are innately predisposed to acquire it will gradually come to predominate in the community. Dennett suggests that this happened with inner verbalization, and I suspect that the same is true of the more complex activity of premising (Dennett 1991a, ch. 7). Normal children spontaneously engage in conscious verbalized reasoning, without instruction or encouragement, and if such reasoning involves premising, as I claim, then it is plausible to think that we are innately disposed to form and execute premising policies. Given the advantages of having a supermind, there should certainly have been the sort of selectional pressure needed for the Baldwin effect to get a grip here. Note that this is not to say that supermental abilities will develop in any environment; some quite specific environmental stimuli may be needed for their emergence – including exposure to and engagement in rational debate, in which rules of inference are followed and at least tacitly acknowledged. But I suspect that in any environment which facilitates acquisition of the component abilities required for premising, children will spontaneously apply these abilities to construct a working supermind. Note, too, that this still leaves room for cultural and individual variation in supermental processes. Even if we are innately disposed to acquire and apply inference rules, *which* rules we acquire will be to some extent culturally determined. Cultural and linguistic factors may also influence the *style* of conscious reasoning – cultures with confrontational debating styles tending to promote more analytical styles of reasoning, and languages that highlight logical form facilitating deductive inference.

This view of the conscious mind may help to shed light on a particularly difficult problem in evolutionary psychology. Much recent work in this area has been inspired by the view that cognition is modular – that the mind consists of discrete modules, each specialized for dealing with a particular domain and with its own dedicated inputs, outputs, and processing mechanisms (see, for example, Barkow et al. 1992; Carruthers and Chamberlain 2000; Hirschfeld and Gelman 1994; Mithen 1996; Pinker 1997). Candidate modules include ones for social contracts, theory of mind, biological classification, and simple physics. This approach has proved fruitful, but leaves us with a problem when it comes to explaining the origins of *domain-general* thinking – thinking which involves uniting ideas from different domains. Clearly such thinking *did* evolve, since we are capable of it, but it is hard to see how a modular mind could support it – there would not even have been a common medium for the construction of cross-modular thoughts. Nor is it clear how a domain-general reasoning system could have evolved from a modular basis; surely any new cognitive demands could have been met more efficiently by tweaking old modules or adding new ones, rather than by developing a cumbersome general-purpose reasoning system? We might call this the 'hard problem' of evolutionary psychology.

Now, if we identify the domain-general system with the supermind, then this hard problem may become at least a *little* more tractable. For on this view, the development of domain-general reasoning would not have required substantial changes to the underlying cognitive architecture. As we have seen, a premising machine could have been created by co-opting pre-existing cognitive resources, with natural language serving as the system's representational medium. Even if a genetic disposition to premising emerged, the neural changes involved need not have been elaborate. Of course, the hard problem immediately resurfaces as the question of how a modular mind could support the various abilities required for premising and how it could co-ordinate them effectively; but refocusing the problem in this way may itself constitute progress. And it is not too hard to see how some of the abilities involved could be supported by modular systems. Meta-representational abilities could be supported by theory of mind or the language system, and strategic skills by social intelligence. Some inferential abilities might also be supported by the language system – taking the form of skills in spotting sentences with certain syntactic features and producing appropriate formal transformations of them. At any rate, there is a potentially fruitful line of research here.[10]

4.3 Clinical psychology

Supermind theory may also shed light on aspects of abnormal development. What would happen if the supermind failed to develop normally? What predictions would the theory make, and are they borne out? Recall the remarks in chapter 5 about the functions of the supermind. There I suggested that supermental abilities confer benefits in the areas of behavioural control, cognitive control, and self-awareness. A person with impaired supermental abilities would have deficits in these areas. They would get stuck in behavioural dead-ends, unable to make up or change their minds; they would have little control over the content or style of their thinking; and they would have a poor understanding of their own minds and motivations.

It is very tempting to make a connection here with *autism*. Autism is a developmental disorder of the brain, characterized by three main areas of impairment: problems in social interaction, communication difficulties, and repetitive behaviour (American Psychiatric Association 1994). It is the last aspect I want to focus on. Autistic people typically exhibit stereotyped movements, insist on following routines, and have a limited range of interests. This aspect of the condition has been relatively neglected in studies. The leading current theory of autism is that sufferers are *mindblind* – that they have an impaired theory of mind system and consequently have great difficulty understanding other people's actions and predicting their behaviour. This theory is supported by some powerful experimental evidence and provides an attractive explanation of autistic people's difficulties in social interaction and

[10] The idea that domain-general reasoning is conducted in natural language and involves the redeployment of existing modular resources is one that has been canvassed in some detail by Peter Carruthers (Carruthers 2002, forthcoming).

communication (see, for example, Baron-Cohen 1995; Baron-Cohen et al. 2000; Leslie 1991). However, it does not provide a direct explanation of their repetitive behaviour, which is often written off as a side-effect of the condition. Since autistic individuals cannot understand and predict other people's behaviour, it is suggested, they find social situations frightening and resort to repetitive behaviour as a coping mechanism (Baron-Cohen 1989; Carruthers 1996a; for criticism see Turner 1997).

In the present context, however, another explanation suggests itself. Perhaps autistic people have difficulty maintaining an effective supermind. This would directly explain their repetitive behaviour. Lacking the power to engage in active premising, they would be at the mercy of their non-conscious minds, unable to override instinctive responses and take active control of their thinking. In contexts where there were clear cues for action this might not matter much. But when confronted with novel situations where creative thinking was required, such people might easily get locked into repetitive behavioural patterns. (There is indeed evidence that autistic people show less repetitive behaviour in situations which are structured for them than in ones where they have to decide for themselves: see Turner 1997.) This hypothesis explains other aspects of autism, too. Autistic people tend to perform well on routine tasks of the sort which can be executed without conscious thought, but much less well on ones that require reflection (see Frith et al. 1994, and the papers in Russell 1997). They also seem to have very impoverished inner lives – as we would expect if they do not engage in supermental activities. When asked to describe their own mental processes they report little or no inner verbalization or unsymbolized thoughts – in strong contrast to normal individuals (Hurlburt et al. 1994; Frith and Happé 1999). Some advocates of the mindblindness theory explain this by claiming that autistic people are blind to their own minds as well as to those of others (Carruthers 1996a), but the present account suggests a more radical explanation – that they simply do not *have* conscious mental states, or only fragmentary ones.

This is of course only a suggestion – no more than a proposal for future work. Let me stress, too, that it is intended, not as an alternative to the mindblindness theory, but as a supplement to it. The mindblindness theory offers, I think, a very plausible explanation of the social and communicative difficulties from which autistic people suffer. But autism is a complex condition, and it is not implausible to suppose that it involves more than one underlying impairment. More intriguingly, it may be that impaired supermental abilities *result from* an underlying impairment in theory of mind. As I mentioned earlier, in order to form premising policies one needs to think about the mental attitudes one has adopted to propositions, and it is very likely that theory of mind is required for this. If so, then a person with impaired theory of mind would have difficulty, not only in understanding the minds of others, but also in constructing a supermind for themselves.

Conclusion

The topics discussed in this chapter by no means exhaust the potential applications for supermind theory, either in philosophy of mind or in scientific psychology. I have, for example, said nothing about how *emotion* might fit into the picture. Do we have two levels of emotion, as well as of cognition? Are the cognitive elements in emotion located at the basic level or the supermental level, or both? How do our emotions affect our premising activities? Supermind theory may also have important implications for theories of mental content. And, of course, the applications already sketched require further elaboration and defence. All this is a matter for another time, however. My main aim here has simply been to set the core theory in a wider context, as a way of helping to evaluate it. I think that the ease with which it finds application in a variety of areas is itself very encouraging.