The Empathic Stance

Examination Number 1860247

MSc by Research

The University of Edinburgh

2010

I have read and understood The University of Edinburgh guidelines on Plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references.

Contents

Chapter 1

Dennett's philosophy of consciousness and two challenges to it

This dissertation takes the philosophy of Daniel Dennett as its starting point. Of course that is far from uncontroversial, but this is not uncritical acceptance. Rather, it is an attempt to discover how Dennett's concepts of consciousness and sentience, in particular, could be modified so as to achieve much wider acceptance.

Introducing The Intentional Stance, Dennett states "I declare my starting point to be the objective, materialistic, third-person world of the physical sciences." (1987a, 5) Much of what he seeks to explain is subjective, however, and so he is committed to explaining subjective phenomena in objective terms. He contrasts this approach with that of Thomas Nagel, who views objectivity, the "view from nowhere" (1986), as incapable of acknowledging the particular points of view of individual sentient creatures. (Nagel, 1979c) In order to clarify *my* starting point, in the first section below I give a short account of the development of my own philosophical views, in which Nagel and Dennett have both played much more prominent roles than any other contemporary philosopher.

Following that initial section this first chapter sets out the Dennettian thesis and two antitheses. Counting Blackmore's argument discussed in

Section 2.2, we have three antitheses, but I believe these are all concerned with the same underlying issue, though they are expressed in quite different terms.

In this chapter, an overview of Daniel Dennett's philosophy of mind, with subsections on the intentional stance, four kinds of mind and the narrative self, is followed by challenges from Andy Clark and myself. Clark questions whether Dennett is right to restrict consciousness to thinkers, or, as Clark puts it, whether "experiences need a thick subject." (Clark, 2002, 197) As I would put it, Dennett fails to explain sentience.

In the next chapter, following a section introducing memetics, Blackmore claims that her experience of "meditative consciousness" (my phrase) cannot be accounted for by Dennett's theory. (Blackmore, 2003)

Chapter 3 begins with a discussion of Clark's (2000) own attempt to solve the problem described in this chapter but finds that (along with some related ideas) wanting, and introduces my concept of empathy as essential to the attribution of sentience. That is followed by a section on the theory-theory, simulation theory and empathy, and than a section on Robbins and Jack's "phenomenal stance" (2006), which is very similar to my own "empathic stance," described in the last section of this chapter.

Chapter 4 is concerned with the relationship between modeling and mentality, concluding that modeling is a better "mark of the mental" than intentionality (Brentano, 1924), and that a minimal, sentient mind is a certain kind of modeler. There is a section on Metzinger's two concepts of selfless consciousness (2004), followed by one discussing self-models and attribution to self and others.

Chapter 5, which is the last, is more speculative than its predecessors. The first section introduces two concepts of information, and uses them to further analyse some of the concepts from the earlier chapters, intentionality in particular. The second and last section reiterates and summarises the main conclusions of the dissertation.

1.1 Background: From Nagel to Dennett

As an undergraduate I became fascinated by the mind/body problem, and was wholly convinced after reading Nagel's famous paper What is it like to be a bat? (1979c) that scientific, third person, objective methods could never fully explain consciousness. Later, I read his essay Subjective and objective (1979b), and, again, found it extremely persuasive. In fact, I'd go so far as to say that Subjective and objective has had a more profound influence on my thinking than any other piece of philosophical writing. My approach to the issues discussed in this dissertation is certainly informed by it.

On the other hand, from my introduction to him in the late seventies (see, for instance, Brainstorms (1979)) through to some point during the nineties, Dennett for me was something of a *bête noire*. I had come to associate the concept of consciousness with humane values, and I felt that Dennett, with the Churchlands (e.g. Neurophilosophy (Churchland, 1989)) and others, through their negative attitude to consciousness, were threatening those values, while Nagel championed them.

From graduation in 1981 to matriculation in 2009 on the course for which this dissertation is written, I retained my philosophical interests, and my views developed. The lasting impact of my reading of Nagel was an appreciation of the significance of subjectivity. I formed the opinion, which I still hold, that, in general terms, subjective and objective phenomena and methods are equally important, though in most specific contexts, one or the other will be more appropriate. For many years I considered myself a dual aspect theorist, and had (as I still do) a great deal of respect for Spinoza. (Nadler, 2007)

Eventually, the concepts of subjectivity and objectivity began to dominate my thinking, so that mental and physical aspects of reality were replaced, for me, by subjective and objective aspects respectively, and I realised that this dichotomy might have a psychological, rather than a metaphysical, basis. But I retained the same sense of the significance of subjectivity, being careful, for instance, to notice and discount any pejorative implication of "subjective" in my reading, and using it myself to mean just "pertaining to or associated with the subject."

A study of Buddhism, some time ago, helped me realise that the concept of the self might lack a referent (Coseru, 2010), and later I began to think of consciousness in a similar way—the concept is useful in some contexts, as is that of the self, but neither need be reified. The values that I associated with consciousness, I realised actually depend on people empathising with each other, i.e. treating each other *as if* conscious, while the *concept* of consciousness, which is what I'd been concerned for, need not be involved. A crucial point was the realisation that we can empathise without ever having thought about consciousness or sentience. So consciousness and empathy changed places, and instead of empathy depending on consciousness, consciousness, or rather the concept of it—because that, now, seemed to be what needed explaining—now seemed to depend on empathy.

These changes in attitude allowed me to return to Dennett, whom I had always found impressive, and now found very convincing. His compatibilism, for instance, I now see as entirely consistent with Spinoza's determinism. I had previously, as a dual aspect theorist, insisted that experience "in its own way" must be just as real as matter, but at the same time I had been dubious about the concept of reality, suspecting that it was often used naïvely, and now I found I could take on board the concept of multiple drafts (Dennett, 1991a), which profoundly conflicts with Nagel's position that the features of a particular experience are a matter of fact (Nagel, 1979c).

The key, I think, to understanding Dennett, is his commitment to "the objective, materialistic, third-person world of the physical sciences." (1987a, 5) Within that limited context, and allowing for the one omission that I seek to remedy in this dissertation, he offers a comprehensive, coherent and, for me, almost entirely convincing account. But he does neglect subjectivity, and especially intersubjectivity, which I find to be central to a full account of consciousness, sentience and their attribution. I have momentarily confused some philosophers, in informal conversation, by saying that Dennett was wrong about sentience signifying mere sensitivity (Section 1.3.2), because it also requires empathy—on the part of the attributor (which of course has implications concerning the nature of the object of attribution). So what attributions of sentience (and, I think, consciousness) really signify is something about the relationship between attributor and attributional ob-

ject. So sentience is not, in my view, an objective attribute, though neither is it merely "in the eye of the beholder," i.e. subjective—rather, it is inter-subjective, concerning *both* beholder/attributor and the attributional object, which is being viewed as another subject.

I believe that Dennett missed the significance of empathy because he was too "objectivist" in his thinking. But I do agree that, *in strictly objective terms*, there's no more to sentience than sensitivity. In fact, something similar can be said about consciousness. That is extremely counter-intuitive because such strict objectivity is so profoundly unnatural in this sort of context that we are hardly capable of achieving it, and of course strict objectivity excludes pro-social (as well as all other) values. It is not that, objectively, we are not conscious: strictly speaking, in objective/scientific/third person terms, the statements that a given entity is *and* is not phenomenally conscious are *both* meaningless. That concept belongs to a different language game (Wittgenstein, 1972), and to try to use it scientifically is a category error (Ryle, 1949).

For me, a newborn child is most certainly the subject of experience—but I don't expect science, or the sort of philosophy that aspires to maximal objectivity, such as Dennett's, to do more than explain why people tend to take that view: what is its psycho-social significance? But Dennett fails, in my view, to do even that. His concept of consciousness confines it to members of a linguistic community (Section 1.2.3). Sentience, for him, is no more than sensitivity (Section 1.3.2). My main aim here is to remedy that failing, by adding what I call "the empathic stance" (Section 3.4) to his array of stances, thus explaining our tendency to see unthinking meditators, members of other species, and newborn children, as much more than mere sensitive mechanisms. But I do, I think, remain close to Dennett on some of the most central points of his philosophy, and the empathic stance is, like the intentional stance, an interpretative strategy: just as there are no intrinsically intentional systems, so there are no intrinsically, objectively appropriate objects of empathy.

1.2 Dennett's theory of mind

1.2.1 The intentional stance

Dennett's *magnum opus* is Consciousness Explained (1991a), and I lean on it quite heavily in what follows, but in the appendix for philosophers (the book is aimed at a general readership) he mentions "the other half of my theory of mind" (Dennett, 1991a, 457), the theory of content that he calls "the intentional stance." This subsection is chiefly concerned with that, but to put it into context, we must also discuss his other two stances. The three stances correspond to three levels of abstraction, of which we will begin with the lowest.

The physical stance is what we take when we think of and deal with a physical object *qua* physical object, such as a rock that might or might not be kicked, depending on its likely mass and the type of footware in use at the time. This stance is probably about as straightforward as it seems, the value of the concept being mainly in contradistinction from the other stances, or just to complete the array (although, as will become evident in Chapter 3, I consider Dennett's array of stances to be incomplete).

The design stance is one step up from the physical stance in terms of abstraction. When you see a tool, such as a hammer, you perceive not just a physical object, but something that was designed for a purpose, and it is quite likely that the concept of hammering, and perhaps that of nails, will appear in your mind, if only at the back. Of course, a hammer can also be dealt with from the physical stance, if, say, you're looking around for something to act as a paper weight when, as in the case of the rock, above, its mass is the main consideration. If you're in the business of doing some hammering, on the other hand, your first resort will be something that you know to have been designed for the purpose, because it is rational to expect that to be the best option, just as a "proper" paperweight will most likely do a better job of weighing down paper than a hammer would—if only because the paperweight will likely be a more convenient size and shape.

Both paperweights and hammers come in a variety of shapes, sizes and materials, and while mass is an important aspect of each, other aspects mat-

ter too. To look for something that was designed for the job that you have in mind is generally more efficient, if such a thing is likely to be available, than to try to select the particular thing, from all of those that are available, whose various properties are most suitable. "Designed for the job" encodes all aspects of suitability: it is a reasonable assumption, though not an infallible one, that the entity concerned will function or "behave" in such a way as to facilitate your purpose. Note, however, that such considerations apply not only to tools: whether we consider living things to have been designed by a deity, or "designed" by evolution, we'll be correct in most cases if we take it that a creature with wings will use them to fly, because that's what they're "for." (Despite the scare quotes, evolution is hereafter assumed to be explanatorily superior to deistic design.)

The intentional stance operates at the highest level of abstraction. "With their wings" is an answer to the question as to *how* winged creatures fly (if not a very informative one), but *why* do they do so? For many reasons, probably, but let's take as an example what's certainly a common one: to find food. In order to understand that—or, at least, as way of doing so that's both extremely common and extremely useful—we view the bird (let's say) as *desiring* food and *believing* that it can be found some distance away—far enough to make flying a better way of getting there—or, perhaps, of spotting it—than walking, hopping, swimming, etc. Belief and desire are intentional concepts, the archetypal ones, in fact, in discussions of intentionality, though there are many others. They are *about* something: desire for food, belief that it might be found somewhere. We might argue about the details of avian psychology, but it would be extremely difficult, while doing so, perhaps even impossible, to dispense altogether with intentionality regarding it, that is, *not* to take the intentional stance toward the bird. Brentano, re-introducing a concept from medieval philosophy, called intentionality "the ineliminable mark of the mental" (Brentano, 1924): all mental phenomena, and no physical phenomena, in his view, are intentional. Dennett's contribution here is to make intentionality a matter of interpretation: to consider something to be an intentional system is to adopt a particular strategy in order to deal with it: to predict future or explain past behaviour.

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. (Dennett, 1987b, 17)

It might seem "natural" to treat birds and other animals as intentional entities—"rational" in the context of avian behaviour does not, of course, have all of the implications that it might when applied to that of a person, but it makes sense in evolutionary terms: the bird's behaviour is rational given its environmental niche and evolutionary goals. Artificial systems, however, are another matter, and to elucidate intentionality in this context Dennett uses the example of a chess playing computer. (Dennett, 1971, 1987b) The attempt to explain and predict the moves of a chess partner is, of course, an essential element of the game. The designer of a chess program, if she were playing against it, might in principle be expected to be able to predict its next move using the design stance, but in practice the program will likely be too complex to allow that, and certainly anyone else, ignorant of the detailed design, would have to resort to the intentional stance: it probably "thinks" it can checkmate me within three moves (and of course it "wants" to do so).

In principle, again, the physical stance could be used to predict what a computer will do, but the details of its operation at that level are far too complex. The design stance, on the other hand, is entirely appropriate for most purposes: if I tap the appropriate key, the letter I want will appear on the screen, because that's how the system is designed. This higher level approach, though it has great benefits, also has a down side: the machine might malfunction and my prediction be thwarted.

It's actually quite hard to imagine using the physical stance in the case of a computer: if I was really using the physical stance, and the appropriate letter failed to appear when a key was struck, that would not surprise me. There is no sense of what's supposed to happen, at that level: that's a design

stance concept. The design stance is intrinsically teleological, and the physical stance not at all so. (In an early paper Dennett says that a malfunction is predictable only from the physical stance (Dennett, 1971, 88), but in fact such a prediction requires switching back and forth between stances.)

But there's a similar discontinuity between design and intentional levels: as an ordinary user of a chess computer, its behaviour in a certain situation might surprise me, so that I suspect it of being faulty. I contact the designer, who tells me, no, that's what it's supposed to do: it's a feature, not a bug. But I realise that the program designer has an imperfect understanding of the game, or perhaps it's my understanding that's at fault, but in any case, in future, whenever that situation arises, I have to drop out of the intentional stance, in which I assume that the computer has the same beliefs, desires, etc. as any other chess player, and descend to the design stance, in which I know that this particular program will do this particular thing, even though it makes no sense to me. To use a higher level of explanation can be very helpful, but it's not a risk-free ride.

What's perhaps most controversial here is that, as already mentioned, for Dennett this is entirely a matter of interpretation—in his view, no system is *intrinsically* intentional: "... a particular thing is an intentional system only in relation to the strategies of someone who is trying to explain and predict its behaviour." (Dennett, 1971, 87) This is undoubtedly a difficult point. In order to adopt the intentional stance toward system B—or, indeed, to adopt any stance at all—system A must be intentional. So if there is no intrinsic intentionality, how can intentional stance-taking get started, how can there be any intentionality whatsoever? I return to this issue in Chapter 4. The intentional stance plays an important part in a number of chapters, though: it was the model for my own "empathic stance," in which the rational agent is replaced by the affective self (Chapters 3 and 4).

1.2.2 Four kinds of mind

Intentionality, for Dennett, extends well "down the scale," encompassing not only all living things, which behave rationally thanks to their design by evolution, but even such simple artifacts as the thermostat, which can be

attributed the belief that the temperature is too hot or cold, and the desire that it be just right (Dennett, 1987b)—even if the usefulness of the intentional stance, as opposed to the design stance, is debatable in this particular case. He divides intentional "creatures" into four classes that correspond to different levels of sophistication.

The lowest class of intentional systems, which would include the thermostat, is **Darwinian**. The thermostat's desired temperature might be adjusted, but its behaviour is basically hard-wired. That's due to evolution in the case of its organic classmates, but for the thermostat it's caused by the activities of creatures that belong to a much higher class and are capable of making and using artifacts (which, to risk getting ahead of ourselves, is also viewable as evolution, but that of memes).

One step up from the Darwinian creature is the **Skinnerian**. As the name implies, these creatures can learn, behavioural strategies being reinforced when rewarded. As well as many species of organism, simple connectionist networks fall into this category. (Dennett, 1997)

Popperian creatures are much more sophisticated still, in that, as well as employing the Darwinian and Skinnerian tricks of their country cousins, they model aspects of their environment, and so can trial potential strategies in a virtual arena, which is more time- and energy-efficient, as well as safer, than the alternative. (The concept of modeling plays an important part in subsequent chapters. Despite Dennett's acknowledgement of it here, in my opinion he generally underestimates its significance.)

The most sophisticated type of intentional system is the **Gregorian** creature. In addition to having the faculties already mentioned, they make and use tools, but not just hammers and the like: the most important tool of all is language, which not only allows much, much more effective communication, but, according to Dennett, programs the individual mind in such a way as to enable great new heights of sophistication to be reached, and in particular, the development of consciousness and the self. So for Dennett, only Gregorian creatures are conscious, all others being mere sensitive mechanisms. I think it reasonable to view this as evidence of Dennett's neglect of affect, which I will argue in later chapters is crucial in the attribution of consciousness or sentience to much simpler creatures. [Important: really

need to return to this point, probably in last chapter.] Meanwhile, however, we continue our survey of Dennett's philosophy, with the development of the self and consciousness.

1.2.3 The narrative self

In The Self as a Center of Narrative Gravity (1992), Dennett adapts Julian Jaynes' (1976) account of the development of "conscious, verbal thought" (Dennett, 1992)[1] as a useful and, indeed, unifying form of "talking to oneself," in which the self is constructed as the hero of the narrative.

Dennett begins his account by describing the concept of a centre of gravity: it is "a purely abstract object", "a theorist's fiction," but "a fiction that has [a] nicely defined, well delineated and well behaved role within physics." (Dennett, 1992)

He goes on to show with examples of its application how "robust and familiar" is the idea of a centre of gravity. Dennett goes to some trouble to counter the negative implications of "abstract object" and "fiction," insisting on the usefulness of the concept. Purportedly causal explanations that feature it can, he says, compete with explanations that clearly are causal. It would be a category error to identify an object's centre of gravity with an atom with which it happened to coincide, it is "...*just* an abstractum," but the centre of gravity is "...a wonderful fictional object, and it has a perfectly legitimate place within serious, sober, *echt* physical science." (Dennett, 1992, emphases in the original)

A self, according to Dennett, is similar: a theorist's fiction that can be extremely useful in characterising the behaviour of things whose activities are relatively complex. He moves on from the centre of gravity analogy to characters in fiction. These are compared to theoretical entities such as atoms, to which, unlike such characters, the "principle of bivalence" does apply. It does not apply, for instance, to Sherlock Holmes: there need be no answer to the question as to whether "he has or had a mole on his left shoulder blade." (Nor to whether the present or past tense is more appropriate when

[1] Page references for the Dennett (1992) paper are not available, because an unpaginated web version was used: see References.

refering to him: whether he is dead or alive.) The self is similar, in that there need not necessarily be an answer to any question that can be asked about one.

But how does a self come into being? Fictional characters require an author, so don't selves, however fictional in themselves, require a real creator of some sort? Dennett thinks not. He develops a scenario in which a novel writing robot creates an autobiography, and therefore a self, out of its own "experiences," despite being, by stipulation, "...not a conscious machine, not a 'thinker.' It is a dumb machine, but it does have the power to write a passable novel." (Dennett, 1992) Dennett perhaps reveals a degree of uncertainty in the potency of this story when he writes

> IF [sic] you think this is strictly impossible I can only challenge you to show why you think this must be so, and invite you [to] read on; in the end you may not have an interest in defending such a precarious impossibility-claim. (Dennett, 1992)

It is tempting to suppose that only those who are already inclined to accept Dennett's account will be receptive to the idea of such a self-generating robot. However, I fall into that category. I find his case, overall, very plausible, if not entirely convincing, particularly so the ontological status of the self, sharing features with both centres of gravity and fictional characters.

Following some discussion of the selves that occur in cases of multiple personality disorder (now called dissociative identity disorder), "normal selves" and fictional characters, in which Dennett continues to stress the similarities (quite plausibly, in my opinion), he goes on to identify our narrative tendencies with conscious thought—he says we are all "inveterate and inventive autobiographical novelists" (Dennett, 1992)—and to offer a speculative evolutionary explanation of it, adapted from Jaynes' account (1976).

First, though, to set the scene, Dennett refers to Michael Gazzaniga's research on "split-brain" patients (he does not provide a reference). He shares the view that he attributes to Gazzaniga: "the unity of normal life is an illusion." Not only are our cortical hemispheres capable of functioning largely independently, if the *corpus callosum* is severed, but "the normal mind is... a problematically yoked-together bundle of partly autonomous systems" that

"sometimes have internal communication problems which they solve by various ingenious and devious routes." (Dennett, 1992)

There is, arguably, some confusion there between personal and sub-personal levels. The unity of consciousness would seem to be one of its defining features, rather than a mere illusion, and it is rather strange to see the mind, as opposed to the mechanisms that underlie it, described as "problematically yoked-together." The question that arises is whether that confusion resides in the theory, or merely in Dennett's presentation of it. I believe that he does veer towards eliminativism at times in his writings, as in this case, where he seems almost to be suggesting that consciousness itself is an illusion, but he also struggles quite hard to avoid it (for instance see his Real Patterns (1991b)), in my view successfully as regards the theory, though not always in the presentation. "Illusion" is probably mischosen: I am sure Dennett would agree that the unity of consciousness is essential to it, while the disunity is entirely sub-personal. The main cause of the confusion is, I think, the fact that, as noted above, Dennett is attempting to explain subjective phenomena in objective terms. Few if any would argue with the suggestion that consciousness is subjectively unified, but Dennett is attempting to delve below appearances here.

Dennett's evolutionary story, briefly, begins with our ancestors being not really conscious, communicating verbally but not consciously. But they could express, or respond to the expression of, a need for help, exchanging information, effectively asking a question or answering it. Due, however, to inter-module communication difficulties within the individual mind, people sometimes found themselves (almost literally) asking a question that they were—in fact, a different module within them was—able to usefully answer, and so the habit of talking to oneself became established, and eventually internalised, as conscious thought.

It has been suggested that Dennett's focus on language leads to an overly intellectualised view (see, for instance, McCarthy (2007)), and that is certainly how I see his work, an example being his tendency to identify consciousness with conscious thought. As we shall see in the following section and Chapter 2.1, at least two other writers (excluding myself) take a similar view, but both of them, like myself, agree with Dennett on the significance

of language, the profoundly transforming effect on the mind that learns to use it.

1.3 Two challenges

This section contains two criticisms of Dennett, and another is described in Chapter 2.1, but in my opinion they all indicate the same omission in his philosophy, which I attempt to remedy in later chapters.

1.3.1 What does it take to have "that special something"?

In a paper entitled "That special something" (reminiscent of Dennett's "factor x" in relation to sentience, Section 1.3.2 below), Andy Clark asks rhetorically "How might the spinning of a narrative (one that cannot be assumed to be already the narrative of a conscious agent) help bring minds like ours into being?" (Clark, 2002)

He mentions four ideas that Dennett has offered in this context, but does not see that any of them "can quite carry the load." These are:

- The idea that linguistic formulation yields a kind of shallow determinacy of content that simple belief-like states lack.
- The idea that linguistically rehearsed contents are especially well-positioned to win the struggle for control of action.
- The idea that any notion of a "point of view" depends on one story winning out over others, and that linguistic judgings are what allow such victories to occur.
- The idea that certain kinds of morally significant self-control require the capacity to confront ones own beliefs and reasons for action, and that the linguistically-supported objectification of our own mental states contributes deeply to this process.

(Clark, 2002) It is not clear to Clark why human experience and "significant suffering" should depend on either of the first two, or why the third is necessary for "pleasure, pain and suffering." (Clark, 2002)

With the fourth, however, "we come closest to seeing some kind of conceptually deep connection between the operation of certain mind-tools and the presence of fully-fledged human agency" (Clark, 2002), and following some discussion, Clark opines that, taken together with other, related aspects of Dennett's philosophy, "as an account of the pre-conditions of morally responsible agency, this... has much to recommend it." (Clark, 2002) Nevertheless, there is nothing in it to account for "the presence of qualitative consciousness and the potential for significant suffering." One further move is required, "to claim that experiences need a thick subject i.e. a subject whose capacities of self-knowledge and self-control lie at, or close to, the apex of deliberative reason," which Clark finds unwarranted. (Clark, 2002)

It seems plausible that Dennett's position that only our judgements about consciousness require explaining leads him sometimes to take what others view as illegitimate short cuts in his thinking. But our intuitions surely also require explanation, whether that results in justification or disillusionment. My own intuitions are entirely with Clark here, and so I set out to show how the attribution of experience, or sentience, to "thin subjects" might be entirely justified.

1.3.2 The missing puzzle piece: sentience?

For Dennett, consciousness develops via language, as we saw in Section 1.2.3. This obviously implies that members of species that do not use language are not conscious. For many people, including myself, this is highly counter-intuitive, until we recall that, by consciousness, Dennett means conscious thought. So we can accept that non-linguistic creatures do not consciously think, or certainly not as we do, while still seeing them (or at least some of them) as being aware of aspects of their environment, suffering pain, enjoying pleasure, and so on—in other words, they are sentient. Unfortunately, Dennett blocks off that avenue too. For him, a sentient creature is no more than a sensitive mechanism. He says that "Everybody agrees that sentience requires sensitivity plus some further as yet unidentified factor x" (1997, 65), but after some discussion he concludes by offering "...a

conservative hypothesis. . . : There is no such extra phenomenon." (1997, 97)

Sentience does not appear in the index of Dennett's magnum opus, Consciousness Explained, and for me that was a glaring omission. The quotes above are from a later work, Kinds of Minds. (1997)

Blackmore's concept of meditation (Chapter 2.1), and Clark's belief that experience does not require such a "thick" subject as Dennett's concept of consciousness implies, both look rather like sentience at first glance, at least. I believe that an explanation of sentience that is more intuitively satisfying while remaining consistent with the main body of Dennett's work is both highly desirable and perfectly possible, and in addition will go some way, at least, towards answering the concerns of Blackmore and Clark. In the chapters to come I develop just such an explanation.

Chapter 2

Memetics and Blackmore's challenge

One function of the first section below is to explain the concept of the meme so that a person initially unfamiliar with it will be capable of appreciating the memetic aspect of my thesis: memetics has an significant part to play in the theories of both Dennett and Susan Blackmore, whose criticism of Dennett's theory of consciousness, discussed in the second section of this chapter, originally inspired this research project. However, the first section also serves another function, to introduce a particular concept of information, which is made use of in Chapter 5.

2.1 Memetics

It seems at least conceivable that, if intelligent aliens could study our culture very closely over a long period without disrupting it—without us having any awareness of their presence—they could eventually discern the shapes of most or even all of our social, political and other institutions, from the family to state regulation of personal finance advisers. These behavioural patterns, though immensely complex and interwoven, are objective: they are what people actually, physically do. This is so even though we feel immersed within them—they are the sea in which we swim—and we naturally take our thoughts and feelings about them, our conscious participation in them,

to be essential to them. But our behavioural patterns are "out there" in the world.

In an attempt to show that not only genes can be considered to evolve, in The Selfish Gene, Richard Dawkins coined the term "meme". (Dawkins, 1976) The word was chosen as reminiscent of "memory," and "gene," and the French word "*même*," meaning "same" ("mimic" has the same roots). It is pronounced "meem". The study of memes is "memetics", on the model of "genetics." "Examples of memes," according to Dawkins, "are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches". (Dawkins, 1976, 192) On this view, such cultural entities quite literally evolve, the meme being to cultural evolution as is the gene to biological evolution. Patterns in human behaviour are viewed as entities that replicate, with variation and selection.

Cultural information, though in some ways analogous to genetic information, consists of items that can be exchanged at any time between members of a society, instead of being transmitted only from parents to offspring at reproduction. Like the gene, the meme can survive the death of all its current carriers, but while genetic information is carried within cells, cultural information reaches the brain via the senses and is stored there, before being transmitted on to another brain, via behaviour (where that includes talking, writing, and other symbolic communication). We do not need to learn everything there is to know about personal finance, but might pick up the idea of consulting an adviser from a friend or colleague, just as the adviser picked up the many, many memes that go to make up her expertise from tutors, books, etc. while training, and some of these memes, but probably not all of them, will be passed on to her clients.

The most obvious type of memetic transmission is where one individual observes another performing some behaviour, and later imitates it. So to start humming a tune you have just heard is a symptom of memetic "infection," as is the picking up of a local accent, and, indeed, learning to speak. The concept of infection is often used in memetics, because unlike genes, which are an essential part of us, memes invade us from "outside." They are the behavioural equivalent of the microbes and other lifeforms that live upon and within us—though of course some of those are beneficial, like the

intestinal bacteria that assist digestion.

But despite the fact that between them, they comprise our culture, not all memes are good for us—most people have experienced a jingle going around and around in the head to the point of annoyance, seemingly beyond conscious control. The fact that we are receptive to items of information does not mean either that we consciously choose to adopt them, or that they must be to our overall benefit. Some political ideas, for instance, as seen by some people as being highly detrimental. The expression "viruses of the mind" is sometimes used for deleterious memes, or groups of them—memes often group together just as do genes, to form "co-adapted meme complexes," or "memeplexes." Examples of types of memeplexes that might be judged beneficial or detrimental are political ideologies and religions. Richard Dawkins is probably best known at the time of writing for his strongly anti-religious views, and seeing religions as viruses of the mind or parasitic memeplexes was an important factor in the development of those views. (Dawkins, 1993) Both Dennett and Blackmore view the self—or at least an aspect or concept of the self—as a memeplex. For Dennett it is a benign "user illusion" (Dennett, 1991a), while for Blackmore it is malign (Blackmore, 2000), though of course such judgements are largely subjective.

What, exactly, is a meme? Most of Dawkins' examples ("tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches") look like objective behavioural patterns, but ideas clearly do not, and he later said that memes are patterns stored in the brain. (Dawkins, 1999) There has been considerable controversy, as to both the value of the meme concept, and also—among those inclined to "believe in it"—exactly what the meme is and where it resides. (Aunger, 2001) For Dennett, memes are "roughly, ideas" (1991a, 201), while Blackmore specifically mentions actions and words but also quotes Dawkins (1976) as above, with apparent approval. (2003, 22)

For reasons that will become apparent in Chapter 5, I have pursued a clearer definition. Historically, the two main competing positions have been "internalism" and "externalism" (Aunger, 2001):

1. the meme is an item of information in the brain, or in the mind—corresponding to and causing the behavioural patterns—or

2. memes are better understood actually to be such behavioural patterns, themselves.

The meme is supposedly analogous to the gene, but unlike that, it has to be transmitted from one location to another without anything like a sperm or egg cell to carry it. The question arises as to how the meme travels, either (1) from brain to brain, or (2) from one instance of behaviour to another—which neither the brain-based nor the behaviour-based school can easily answer. In each case there is a gap that appears unbridgeable.

It seems to me that the best case that can be made for memes requires viewing them as items of information that are encoded in both brains and behaviour, travelling between brains via behaviour, and between instances of behaviour via brains. It might be helpful if I make it clear at this point that I consider the meme to be an objective entity, and that I am currently concerned only with objective aspects of memetics. I relax that constraint later in this section. ("Encoded," in this context, and the ontological status of items of information, are discussed in Section 5.1.)

On this account, memes have two phases of existence: encoded in the brain, and encoded in behaviour. A brain that carries a particular meme is motivated, in appropriate circumstances,[1] to perform the relevant behaviour—the circumstances comprise the decoding mechanism. Another brain, observing that behaviour, processes the incoming information with the result that the meme takes up residence in this brain—is encoded there—and if the action is again triggered and observed, the cycle continues, brain to behaviour to brain, and so on, each transformation being viewable as encoding from one point of view and decoding from the other. The pattern within the brain's physical information that corresponds to a given meme, is the behavioural pattern, encoded, and the behavioural pattern is the neural pattern, encoded.

But there's an apparent difference between this and other en/decoding

[1] "Circumstances," here, should be read as not just the external environment, but also the "internal environment," that is, other aspects of body and brain.

cycles, such as between sentences in different languages. It might be thought that, while the two sentences in different languages have in common their meaning, the brain and behavioural encodings have nothing in common beyond the fact that each one is caused by the other (in the sense that chicken and egg are caused by each other), so that "meme" in this scenario does not actually refer to anything. The two sentences mean the same, but the two meme-forms "mean" each other, and the scare-quotes are significant, because this sense of "meaning" is unusual, to say the least.

The solution to this problem is that, in fact, there's no difference between brain and behavioural patterns

1. as each other, encoded, or
2. as different encodings of a single abstraction.

These are isomorphs: they are computationally equivalent. But for our convenience, 2 is preferable—this abstraction is what we call the meme. It is what an instance of behaviour and the neural pattern that corresponds to it have in common. The gene is a function of the relationship between DNA and its cellular context (Dennett, 1995a), and the meme is similarly abstract. However (relaxing the objectivity constraint), the abstraction can be conveniently encapsulated by saying that the meaning of the meme, what the different encodings have in common, is the significance in that culture of the corresponding action, whether that's humming a tune or throwing a pot. That meaning or significance is the intersubjective aspect of the meme.

So, do memes "really" exist? For me, the reality of repeated patterns of human behaviour, picked up by one individual and passed on to others, is indisputable—and that's really all that's required. Acceptance of that implies acceptance of the whole brain-to-behaviour-to-brain-to-behaviour information en/decoding story, as there is really no other way for such patterns to propagate. I suspect that some of those who find the proposition to be much more controversial are confusing memetics with related issues such as: is the concept of any real benefit in understanding particular aspects of particular cultures? Or, what would the acceptance of it imply for our attitudes towards individual responsibility? Was it me, or was it my memes that did it? That's a serious issue, and quite close to the main concerns of

this dissertation, but that's no reason to let it muddy the water—objective behavioural and neural patterns do exist, and cultural information is transmitted by them "horizontally" through human society, just as genetic information is transmitted "vertically" by reproduction.

It seems certain that simple imitation of behaviour, modeling or mimickry, was the original method of memetic transmission (Blackmore, 1999), but with increasingly sophisticated human communications, they have found shortcuts from one brain to another. A particular type of behaviour might spread, not just by being directly copied, but by being talked about. Replication still takes place, but it is not due solely to direct imitation—we often do things as a result of being told about them. And as well as talking, of course, we have writing, and all of the more recently developed media. There is an important distinctions to be drawn between symbolic and non-symbolic communication. With direct imitation, the meme is encoded in behaviour (as well as brains), but with symbolic communication, it is encoded in words (as well as numbers, etc.), whether written or spoken. Now, communicating, in any way, is certainly a form of behaviour, but this is another level of encoding, beyond basic neural and behavioural encoding.

2.2 Blackmore's challenge

In a special issue of the Journal of Consciousness Studies devoted to Machine Consciousness, Susan Blackmore questions Daniel Dennett's concept of consciousness. She quotes Dennett: "Human consciousness is *itself* a huge complex of memes. . . " (Dennett, 1991a, 210), goes on to describe her experience and understanding of meditation, and concludes:

> *If* this experience can justifiably be thought of as consciousness without memes, then there is something left when the memes are gone, and Dennett is wrong that consciousness *is* the memes. (Blackmore, 2003, 25, emphases in the original)

It seems possible that these are different concepts of consciousness, that is, different explananda, but Blackmore appears broadly to agree with Dennett on the nature of "normal" human consciousness; her discussion of ma-

chine consciousness concerns that rather than meditative consciousness and seeks to explain the possible occurrence of the consciousness meme in machines; and she sees normal human consciousness as illusory: not that it does not really exist, but "it is not what it appears to be." (Blackmore, 2003, 22) However, she believes that in meditation, memes can be temporarily eliminated, resulting in a state of non-illusory consciousness, by which she means that there is no longer a sense of a self that is conscious, no "user illusion," which she sees as malign (Blackmore, 2000), unlike Dennett, for whom it is benign (Dennett, 1991a).

As already mentioned, Dennett does generally lean very heavily on language, and for him memes are primarily linguistic. He says that they "... are, roughly, ideas," and he gives a list of examples that arguably contains few, if any, that could be communicated, or even in some cases implemented, without language:

> the ideas of wheel, wearing clothes, vendetta, right triangle, alphabet, calendar, the Odyssey, calculus, chess, perspective drawing, evolution by natural selection, Impressionism, "Greensleeves," deconstructionism. (1991a, 201)

In connection with the potential immortality of memes he mentions books and inscriptions on monuments.[2] (1991a, 205) Another list of examples is very heavily laden with those that, as in the initial list, require language:

> ...cooperation, music, writing, environmental awareness, arms reduction... The Marriage of Figaro, Moby-Dick, returnable bottles, the SALT agreements... shopping malls, fast food, advertising on television... anti-Semitism, hijacking airliners, computer viruses, spray-paint graffiti." (1991a, 203)

This is not to say that, for Dennett, memes are exclusively linguistic—music is an obvious exception—just that they are predominantly so. Taking that

[2]He cites these not as examples of potentially immortal vehicles, but as artifacts that might seem relatively permanent but nevertheless have limited life spans, saying that "As with genes, immortality is more a matter of replication than of... longevity..." (1991a, 205)

with his comment on consciousness and memes quoted in the first paragraph of this section, it seems not unreasonable to suggest that the narrative self could also be termed the memetic self. In other writings Blackmore develops a concept of the self as a co-adapted meme complex, or memeplex, going so far as to coin the term "selfplex" to denote the relatively special nature of this particular memeplex. (Blackmore, 1999) The valuation of the self and the issue regarding the memetic nature of consciousness are, according to Blackmore, the two main differences between Dennett and herself in this area. (Blackmore, 2003, 25)

However, on the matter of consciousness she might be accused of selective quotation, because following directly on from "Human consciousness is *itself* a huge complex of memes," the original continues, "(or more exactly, meme-effects in brains)..." (Dennett, 1991a, 210), which makes her case rather less clear-cut. Might not the effects remain when the memes themselves have been temporarily eliminated?

Dennett also has another possible line of counterattack: in Consciousness Explained, from which that quotation is taken, he is concerned with normal human consciousness, not altered states, and it might therefore be unreasonable to expect his theory to cover them. For these reasons, although I do believe that Blackmore is on to something, I recruited Clark, as already discussed in Section 1.3.1, to back up my own concerns regarding sentience (Section 1.3.2). Memetics does not play a crucial part in the arguments of the succeeding chapters, but it plays a part in Chapter 5.

Chapter 3

Phenomenality and Empathy

In the first section of this chapter I look at some arguments that attempt to derive phenomenality from functionality, criticise that approach, and suggest instead a link between empathy and the attribution of consciousness.

The second section provides a discussion of empathy, relating it to theory and simulation theories of folk psychology, and the third is concerned with the "phenomenal stance," which is very similar to my own "empathic stance," described in Section 3.4.

3.1 Arguments from functionality to phenomenality

In A Case Where Access Implies Qualia? (Clark, 2000), introducing his own "attempt to complete the puzzle" described in Chapter 1 (the phrase is from Clark (2002, 197n)), Andy Clark questions Block's (1997) distinction between access and phenomenal consciousness. According to Block, we have access to the contents of consciousness when they are available for action control, verbal report or reasoning, but the phenomenal aspect of consciousness, what it actually feels like to experience something—or the fact that it does feel like something to experience anything—is another issue. To explain the latter is Chalmers' "hard problem" (1995), and Block and Chalmers agree that many of the purported explanations of phenomenal consciousness actually address access consciousness, or aspects of it, i.e. some of Chalmers' relatively "easy problems," while these two concepts are really

quite distinct. (Clark, 2000, 30)

Clark, however, pace Block and Chalmers, argues that there is at least one case in which the facts regarding access imply phenomenal consciousness, indeed that given these facts "it is impossible to conceive of the absence of phenomenal consciousness." (Clark, 2000, 30) I disagree with that statement, as do Myin and O'Regan (2002), but I also disagree with their alternative suggestion, and for the same reason. This section is concerned with Clark's and Myin and O'Regan's accounts, and briefly states my disagreement with them, which relates to empathy.

Clark asks us to consider a hypothetical system, capable of distinguishing differences between stimuli, that can be interrogated. If it detects, say, a difference in colour between two stimuli and is asked about that detection, then, Clark suggests, it must reply **either** that it has no access to the act of detection, and just knows about the difference without any knowledge of how it has come to know, **or** that it has access to certain aspects of the detection process, and is "non-inferentially aware" (Clark, 2000, 30), for instance that the sensory modality involved is visual rather than tactile. In the latter case, according to Clark, "it must say that there is something it is like to see the difference rather than e.g. to smell it," (Clark, 2000, 30) and so a fact about access implies phenomenality.

Clark's story, in his view, "shares much of the flavour" (Clark, 2000, 31) of Dennett's explanation(s) of consciousness, while focusing on a specific type of report and access rather than broad availability. But do we need to explain qualia, or merely to explain why we judge there to be such? Describing Dennett as a prime example of the latter camp, Clark suggests that we "need to do more—but not much more!" (Clark, 2000, 32) What more we need to do is "to show how these sincere judgements [that there are qualia] can be true." (Clark, 2000, 32)

Clark claims that "...honest reports of *genuine*, direct, non-inferential access to acts of perceptual difference detection imply the presence of genuine *phenomenal* differences" (Clark, 2000, 33, emphasis in the original), the alternative being "strictly inconceivable," so he identifies such access with the existence of qualia. And this is verifiable, at least in principle: "There is a pattern of actual (ultimately neurophysiological) access which is

underwriting the judgements." (2000, 33)

As Clark acknowledges, despite the possibility of verification, it could still seem conceivable to some that a zombie, utilising the same neural "machinery," could make the same reports and yet be experiencing nothing whatsoever. Clark's position is reminiscent of that of Dennett on zombies (Dennett, 1995b), saying that such conceivability is illusory, because in reality this is "strictly inconceivable." (2000, 33) The issue of the conceivability of the philosophical zombie (Kirk, 2006) comes up again later in this section, and again in later chapters.

In a later paper, Ward, Roberts, and Clark (forthcoming) return to these issues. They summarise the previous paper thus: "Clark argues that certain patterns of access-consciousness... actually entail phenomenal consciousness," and as for the main topic of the later paper, they "see the action-space account as an empirically-motivated way of further fleshing out this proposal." (Ward et al., forthcoming)

In response to the hypothetical challenge that the action-space account explains merely the propensity to judge that there are phenomenal states, they write "this objection stems from a mistaken conception of experience." (Ward et al., forthcoming) They cite Chalmers (2004) as distinguishing between two camps regarding the relationship between consciousness and intentionality. One camp, that includes, for instance, Dretske and Tye,

> attempts to ground consciousness in intentionality, and to do so "without remainder": that is, they argue that there is no more to various states of conscious experience than the obtaining of various intentional and content-bearing representational states. (Ward et al., forthcoming)

The other camp attempt to ground intentionality in consciousness. "The action-space account" according to Ward et al. "belongs firmly in the first of these two camps." Instead of seeing properties of experience as instantiated by the experience, we should view them as "intentional properties... of how that experience represents the world as being." (Ward et al., forthcoming) (This line of thinking follows Jackson (2003).) To view experience as instantiating properties is, at least arguably, to reify experience, and to argue

against experience as instantiating properties is to argue against qualia as properties of experience, as these authors realise:

> To be apprised of one's poise over an action space is to know what one can do. Blindsight, and certain other pathologies of conscious experience, thus emerge as failures of knowledge and representation, rather than as failures to be acquainted with mysterious "qualia". (Ward et al., forthcoming)

Of course, on a purely subjective reading of "qualia," there is no controversy, but that appears not to be what Ward et al. have in mind here. (See Tye (2007) on differing concepts of qualia.) This approach seems to me to be heading in the right direction, but it does not go far enough, eliminating (a less than purely subjective concept of) qualia but retaining a functionalist account of phenomenality. Phenomenality comes up again in Section 3.3, and later chapters.

In the earlier paper, Clark adds to the argument from inconceivability a "skill-theory" of perception that is consistent with his take on access and phenomenality, and in the later paper that skill-theory develops into the action-space account, which I find highly convincing, as an account of the functionality associated with human consciousness. But, as in the earlier paper, we are asked to believe that any system demonstrating such functionality should be considered phenomenally conscious, with all that that implies (see Section 3.3 and later chapters).

Myin and O'Regan (2002), referring to the earlier paper (Clark, 2000), write

> . . . it seems to us that a robot that would be able to discriminate whether an input was visual and not tactile would not *thereby* become a better candidate for being called conscious than a robot that would be able to discriminate whether an input was red versus green. . . Similarly, we don't see why someone who believes neurocomputational mechanisms can never account for phenomenal consciousness would have to change their opinion in the case where the neurocomputational mechanisms mediate

> access to modality, rather than access to sensory [quality]. (Myin and O'Regan, 2002, 29, emphasis in the original.)

But that line of criticism applies equally to any argument that attempts to derive phenomenality from functionality, including that of Ward et al., and Myin and O'Regan's own (see below). Functionality, as a set of input-output relationships, is observable by a third person, but phenomenal consciousness by definition is not, so skepticism regarding phenomenality is irrefutable, as we see in the problem of other minds. Reasoning regarding it therefore tends to lean heavily on intuition. For Clark a system that can verifiably report not merely what it perceives, but an aspect of the nature of that perception, simply must be phenomenally conscious, whereas Myin and O'Regan just do not see that access to modality, however reliable, necessarily implies that there is something that it is like to be the entity concerned.

It was of course Nagel who famously stated that an entity is conscious *iff* there is something that it is like to be that entity. (1979c) At this point I will introduce my own view, to be developed in subsequent chapters: as I see it the primary difference between Clark on one side and Myin and O'Regan on the other is that Clark feels compelled by his consideration of this particular type of access to imagine himself in the place of the system concerned, to project aspects of his own experience onto it, to empathise with it, while Myin and O'Regan feel no such inclination in this case. So my position is that to try to imagine what it is like to be a particular entity is to seek to empathise with it, and so to believe that there is something that it is like to be some thing is to believe that it is the sort of thing with which empathy is possible. On that basis our judgements about qualia and phenomenality will, I believe, be entirely explained, which is all that is required within an objective, third-person, scientific conceptual framework. So, unlike Nagel, I do not view either what it is like to be a particular thing at a particular time, or whether it is like anything, as a matter of fact—or not if "fact" implies objectivity. What it is like to be some thing, and whether it is like anything, for me are not objective issues but intersubjective ones. And that has positive implications: in my view, it is most definitely not wrong to view the sort of system discussed by Clark as instantiating phenomenality. What

would be wrong, however, is to claim that such a system, or any system, *really, objectively, does* experience anything, and it is also wrong to do as Clark does, and insist that we must view it as doing so.

Though not persuaded to attribute phenomenality to the system under consideration by Clark's arguments about access, Myin and O'Regan take his suggestion as to how sensory modalities might be distinguished—in terms of the different skill sets afforded by different modalities—and place that at the centre of their attempt to "naturalize phenomenology." (2002, 27) Citing Evans (1985) and Grush (1998), Clark suggests that "what marks information as belonging to one modality rather than another is the way it is positioned to guide skilled activity." (Clark, 2000, 34) For instance, visual information in many cases will facilitate reaching out to grasp an object, while no auditory information would allow such an action, or at least facilitate it to the same extent. This is the "skill theory" mentioned above.

Having rejected Clark's account of phenomenality, Myin and O'Regan begin their own analysis by asking, rhetorically, "what is it like to see?" (2002, 30) From traditional analyses they derive a number of properties that they take to be representative: ongoingness, forcible presence, ineffability and subjectivity—though this list is not claimed to be exhaustive.

For them, again, phenomenality is a matter of functionality: just as life is not a vital spirit possessed by an organism, but a set of capacities such as movement, respiration and reproduction, so

> . . . phenomenality is not caused by some brain process, but is constituted by the different *capacities* that "feeling" involves. . . And each of these capacities, since it is functionally defined as a *capacity*, must naturally have a functionally describable, and so scientifically amenable explanation. (2002, 33, emphasis in the original.)

They go on to offer a specific skill theory that they claim accounts for the previously listed properties of phenomenality, but we need not concern ourselves with the details. What the authors of the three papers considered in this section have in common, in my view, is that they describe systems onto

which they think any reasonable person should project the concept of qualia or phenomenality—but if such a a system (or any system) is fully described in functional, scientific, third-person terms, it will be possible to view it as "mere" mechanism without compromising one's understanding of its operational principles. That might well hamper one's ability to explain or predict particular behaviours, "on the ground" as it were, but that is entirely due to our own limitations.

Certainly, the more sophisticated a system, and the greater its apparent similarity to ourselves, the more plausible will seem the claim that there is something that it is like to be that system—as I think is clearly demonstrated by the action-space account of Ward et al. (forthcoming)—but I think that claim ultimately concerns not facts, but values: not whether it "really" possesses phenomenality, but whether we should deal with it as if it did, which for me means whether we should empathise with it.

Consider the popularity of the concept of the philosophical zombie. Dennett says that we can't "really" imagine a creature that seems, to all appearances, to be a perfectly normal human, while in fact having no phenomenal consciousness whatsoever. (Dennett, 1995b) But it certainly seems to many of us that we can, indeed, imagine just such a thing. I believe that the philosophical zombie is just an imaginary person with whom we imagine having absolutely no empathy whatsoever. The absence of empathy—if complete, at least—is psychologically equivalent to the denial of sentience/consciousness.

3.2 Theory-theory, simulation theory and empathy

For over two decades there have been two main competing theories of folk psychology, but relatively recent research seems to indicate that, properly understood, they are both valid, and indeed complementary. Their present relevance lies in the fact that they correspond to two forms of empathy, one of which, affective empathy, is neglected by Dennett, to the detriment of his philosophy of consciousness, in my opinion.

In general terms, a substantial proportion of social interaction consists of, or relies upon, the attempted prediction by one person of another's future

actions, or explaining past actions. It is, perhaps, a necessary component of your belief that you "know someone" that you think you can successfully predict how they would act in some situations. And, of course, there are some situations in which we think all normal people would behave similarly: finding themselves in danger, for instance, they would seek to escape from or mitigate that danger. So people generally have some understanding of each other, or at least think that they do, and this supposed understanding has commonly been called "folk psychology."

The main alternative theories of folk psychology are the "theory-theory" and the "simulation theory." The former takes folk psychology to be a quasi-scientific activity, involving entities such as beliefs and desires and hypotheses about law-like regularities governing their interactions. The alternative theory is that we "simulate" other people's minds, in order to predict and explain their behaviour, imaginatively putting ourselves in their situation. Hybrids of theory-theory and simulation theory have also been suggested. The first subsection below is concerned with theory-theory, the second with simulation theory, and the third with hybrids and empathy.

3.2.1 Theory-theory

The term "theory of mind" was first used in psychology by Premack and Woodruff (1978). Their work with a chimpanzee named Sarah seemed to demonstrate that she attributed desires to a human pictured in certain dilemmas, which she "solved" for him by selecting pictures depicting the appropriate action on his part. The authors suggested that the system of inferences implied by Sarah's choices "... is properly viewed as a theory." (1978, 515) Premack and Woodruff's theory accordingly became known as the "theory-theory."

This implies that Sarah adopted the intentional stance towards the human, but in a simultaneously published peer commentary, Dennett (1978) argued that theory of mind is very difficult or impossible to demonstrate unambiguously where language cannot be used to communicate with the subject. (Regarding the significance of language in Dennett's philosophy, see Section 1.2.3 and Chapter 2.) Investigations into the development of

folk psychology in children followed. Wimmer and Perner (1983) (the title of whose article echoed that of Dennett's commentary) reported that the three-year-olds they tested appeared not to understand that a person might hold false beliefs. Baron-Cohen et al. (1985) designed a different version of that experiment, which has since been repeated with various modifications, being generally referred to as "the Sally-Anne test," after the characters in the scenario, or "the false-belief test."

Sally and Anne are two dolls. Sally has a basket, and Anne, a box. The child is shown Sally putting a marble in her basket and then leaving the room. While Sally is out, Anne takes the marble from Sally's basket and puts it in her own box. Sally then returns, and the child is asked where Sally will look for her marble. The correct answer is "in the basket," but very young children appear to have no concept of false belief, and say "in the box," presumably because that is where they know the marble is. Older children, however, from around four years of age, generally "pass" the test (Baron-Cohen et al., 1985): they correctly attribute a false belief to Sally, having, according to the theory-theory, become able to recognise that other people have minds of their own. This is the concept of the child as "little scientist." (Gopnik and Meltzoff, 1997)

Perhaps the two most obvious questions for theory-theorists concern the relatively early age at which children start to think about other people's minds as such, and the ability, or lack thereof, of most people to discuss (folk) psychological theories. However, Davies and Stone (2001), following Goldman (1989), make a comparison with language acquisition, pointing out that

> . . . almost nobody now thinks that there are good objections to the whole enterprise of Chomskyan linguistics starting from the fact that ordinary folk are not very good at articulating grammatical principles. Nor is linguistics threatened by a problem about early acquisition. The linguist can respond to the two putative objections by saying, first, that knowledge of language is partly tacit and, second, that it is partly innate. (Davies and Stone, 2001, 21)

And the theory-theorist can respond similarly: if our folk psychology is partly tacit and partly innate, then we should not be surprised that evidence of it appears at a relatively early age, and that people generally would encounter difficulties in attempting fully to articulate it.

3.2.2 Simulation theory

A rival explanation emerged in the mid-eighties. Heal (1986) and Gordon (1986) independently suggested that, rather than theorising about Sally's mental state, the child would imagine herself in Sally's position, or "put herself in Sally's shoes," enabling her to *simulate* Sally's thinking and behaviour.

> The basic idea is that if the resources our own brain uses to guide our own behavior can be modified to work as representations of other people, then we have no need to store general information about what makes people tick: We just do the ticking for them. (Gordon, 2009)

Heal (1998) later argued that simulation can be understood in two different ways. It was generally assumed that the theory-theory versus simulation theory question is an empirical one (Boden, 2006), but simulation theory is necessarily true, albeit in a weak sense, if, in wondering whether someone likes coffee, I necessarily think about coffee. Strictly speaking, a person using pure theory could think only about preferences, beliefs, etc., and not about coffee. If both the person whose tastes I am considering, and I myself, are thinking about coffee, we are "co-cognizing," in Heal's terminology, which amounts to simulation on my part. If we take the view that thinking about thinking about coffee necessarily implies thinking about coffee, theory-theory would appear to be a priori false, and simulation theory a priori true. (Heal, 1998)

This might lead you to conclude, with Boden, that "[t]he empirical question is what 'sub-personal cognitive machinery' is involved in implementing such co-cognition." (Boden, 2006, 489) However, Heal argues that such thinking leads to a "threat of collapse:" if, in simulating another's thinking, we use mechanisms that are (substantially) the same as their's, then surely

tacit knowledge of others' minds is embedded in one's own mind, and simulation reduces to theorising. (Heal, 1998)

Davies and Stone (2001) argue that this threat is illusory, saying that knowledge implies representation, and, where one mechanism is used to simulate a similar mechanism, no representation is directly involved. The assumption that the mechanisms are sufficiently similar—that I am so like you, say, that I have a good chance of guessing correctly what you might do in a given situation by imagining myself in your place—does seem to require representation, but that is part of the minimal theory that any simulationist is bound to accept anyway. (Davies and Stone, 2001) [If the tendency to simulate others is innate, as some writers suggest (Davies and Stone, 2001) and I believe, then perhaps even such minimal theory need not necessarily be involved in every instance of simulation. I will return to this issue later, either in a subsequent chapter or by rewriting this section.]

If we allow that there is, after all, no real threat of empirical, subpersonal simulation collapsing into theory-theory, is there any actual evidence for such a mechanism? Gallese and Goldman (1998) use findings concerning "mirror neurons" in macaque monkeys to support an argument for simulation theory. Mirror neurons are so named because they fire both when the animal makes a certain action and when it observes another performing the same action. Individual mirror neurons have not been found in humans (there are problems with recording single cell activation in humans), but studies using various neuroimaging techniques have revealed "mirror systems" that include the areas in which mirror neurons have been found in monkeys, but also others including the somatosensory cortex, which, it has been suggested, allow people to know what it feels like to perform the observed action. (Gazzola and Keysers, 2009)

Saxe adapts the "argument from error" of Nichols and Stich (1995) "to show that the errors that human observers make are not consistent with the 'resonance' simulation theory embraced by mirror neuron enthusiasts. Rather, observers must rely on a naïve theory of psychology." (Saxe, 2005a, 174f) Mitchell (2005) takes the target of Saxe's argument to be simulation theory generally but that is not the case, in fact Saxe explicitly promotes simulation/theory hybridism, as we shall see in the next section. Nor does

she reject altogether the mirror arguments: "The mirror system does offer powerful insights into the neural representation of simple actions and some basic emotions. . . " (Saxe, 2005a, 174)

3.2.3 Hybrid theories and empathy

Theodor Lipps, writing in German on aesthetic appreciation (1903), coined the term *Einfühlung*, literally "feeling-into," which came to be translated as "empathy." (Wispé, 1990) He believed that it is through empathy that we come to know others: we do not perceive such emotions as pride, shame, anger, sorrow, and joy in others directly, but experience them vicariously. We "feel for" the other person. (Wispé, 1990)

The psychologist G.W. Allport disagreed. He argued that knowledge about others must be something more than empathy.

> A "proud" gesture, a "joyous" laugh, describe those qualities in another sentient being. So first there must be a realisation of the consciousness of the other. There can be no proud or joyous *stones*. (Allport, 1937)

The crux of that claim is the "realisation" of the other's consciousness. It would seem to imply that, in every case in which we might experience empathy, we first must go through the experience of saying to ourselves, in effect, "this is a conscious being." It seems to me, on the basis of ordinary experience, that this is simply not what happens. On the other hand—at the other extreme—if it means that, in human psycho-social development, children must realise that others are conscious before they ever feel empathy, then that seems equally wrong.

Logic might seem to suggest that belief in the consciousness of another must precede any belief regarding the content of that consciousness, but we are not primarily logical creatures. In *psycho*logical terms, there is no reason why sharing of emotion should not come first, and rationalisation regarding the status of the empathic stimulus, later. Allport has it back-to-front: he thinks that he must know that the other is conscious before he can empathise with them, but in fact he only wishes to attribute consciousness

because he tends to empathise. [Should this stuff be held back to go with attribution rather than empathy?]

Despite its relatively recent coinage, there are now many different definitions of empathy; Batson (2009), for instance, describes eight that he has found in the literature. However, he views these as "related but distinct phenomena," and sees that range of definitions as resulting from attempts to answer two different questions (Batson, 2009, 3):

1. "How can one know what another person is thinking and feeling?" and

2. "What leads one person to respond with sensitivity and care to the suffering of another?"

The main concepts of empathy as described by Batson are (2009, 4ff):

1. Knowing another person's inner state, including his or her thoughts and feelings;

2. Adopting the posture or matching the neural responses of an observed other;

3. Coming to feel as another person feels;

4. Intuiting or projecting oneself into another's situation;

5. Imagining how another is thinking and feeling;

6. Imagining how one would think and feel in the other's place;

7. Feeling distress at witnessing another's suffering; and

8. Feeling for another person who is suffering.

Concept 1 corresponds most closely to the first of Batson's two questions, and he found that, despite all eight concepts having individually been classified as empathy, each of the five concepts 2–6 has been invoked to explain how the situation expressed by concept 1 can come about. Batson suggests, seemingly somewhat arbitrarily, that the theory-theory might be used to explain how concept 1 is realised via concept 5, and simulation theory used

for concepts 3, 4 and 6. He exempts concepts 7 and 8 from being considered in the context of theory- and simulation theories because they "... are not sources of knowledge (or belief) about another's state; they are reactions to this knowledge. Thus, they are not likely to be invoked to explain how one knows what another is thinking and feeling." (Batson, 2009, 9) Instead, they are relevant to question 2.

Batson's account aligns with a view common among theory/simulation hybrid theorists, that we use theory in some situations and simulation in others. (Saxe, 2005b; Gordon, 2009) Saxe, however, considers these accounts unsatisfactory. In reply to Mitchell (2005) she writes:

> Proposals for how to distinguish the contexts requiring simulation or theorizing seem unnatural, for example dividing brief (simulation) from longer-term mental states (Perner and Kuhlberger, 2005), or accurate (simulation) from inaccurate attributions (Nichols and Stich, 2003)... More importantly, in these models simulation and theorizing exist side-by-side but independently, and the observer uses them one at a time. If anything, the dichotomy between the two processes is enhanced... Rather than focus on the circumstances in which observers either simulate or theorize, I prefer to ask how the separate intuitions that motivate [simulation theory] and [theory-theory] can be integrated into a single more general model. For example, how could a naïve theory of mind be informed by the observer's own experiences? (Saxe, 2005b, p. 364)

Shamay-Tsoori (2009) reviews the literature on empathy and, like Batson, finds a number of different concepts, but she divides them into two categories: "The critical difference between cognitive empathy and affective or emotional empathy is that the former involves cognitive understanding of the other person's point of view whereas the latter also includes sharing of those feelings..." (Shamay-Tsoori, 2009, 215) The investigators who focus on affective empathy "typically study aspects such as helping behavior" (Shamay-Tsoori, 2009, 215), so Batson's concepts 7 and 8 fall into this category.

Following brief descriptions of theory-theory and simulation theory, in connection with which she mentions mirror neurons, Shamay-Tsoori goes on to suggest that the former

> . . . views empathy as a thoroughly "detached" theoretical analysis that involves areas of the cortex that are usually activated during mental state attribution, whereas simulation depicts empathy as incorporating an attempt to replicate the other's affective mental state via neural networks related to emotion processing. . . it may be suggested that cognitive empathy involves more [theory] processing, whereas affective empathy involves more simulation processing. (Shamay-Tsoori, 2009, 216)

Watson and Greenberg (2009) explicate what Shamay-Tsoori implies there: mirror systems reflect emotions as well as actions. "Regions in the brain associated with feeling a specific emotion are activated by seeing that emotion in another or witnessing the other in a situation that might elicit the emotion." (Watson and Greenberg, 2009, 126) More specifically, Pfeifer and Dapretto (2009) report, in children asked to imitate or just observe various emotional facial expressions, "significant correlations between activity in mirror neuron and limbic regions and each of the first three subscales of the Interpersonal Reactivity Index" (Pfeifer and Dapretto, 2009, 188), a widely used test of empathy, of which these subscales measure the more affective components.

Affective empathy plays rather an important role in my thinking, and there is one definition of it in the literature, by de Vignemont and Singer, that fully accords with my own view:

> There is empathy if: (i) one is in an affective state; (ii) this state is isomorphic to another person's affective state; (iii) this state is elicited by the observation or imagination of another person's affective state; (iv) one knows that the other person is the source of one's own affective state. (de Vignemont and Singer, 2006, 435)

It seems clear to me that the categorisation of empathy into cognitive and affective varieties, and the close association of these with theory-theory

and simulation theory respectively, are quite solid findings: in the broadest of terms, we have two ways of relating to each other, of which Dennett's array of stances acknowledges only one. The next section discusses one attempt to remedy that situation by adding another stance to the array, but finds it wanting, and the final section in this chapter describes my own candidate for that position.

3.3 The phenomenal stance

Robbins and Jack begin their article "The phenomenal stance" (2006) with a discussion of the intuition that underlies the explanatory gap between the physical and the personal. Following Levine (1983) they divide philosophers who have written on this into two camps: those who take the intuition seriously and endorse some form of dualism, such as Chalmers (1996), and those who reject the "special difficulty" as illusory, like Dennett (1991a).

Robbins and Jack agree with the rejectionists that the intuition behind the explanatory gap should not be taken at face value, but they do take it seriously, seeing it as "psychologically real and deep," itself requiring explanation. In this paper they claim to offer "an empirically grounded account of how the intuition arises." (2006, 60)

Dennett's physical and intentional stances are "mapped" by Robbins and Jack onto folk physics and what they call "mindreading," respectively. They identify mindreading with theory of mind (Section 3.2), and say the correspondence with the intentional stance is only "loose and approximate" but is sufficiently close for their purposes. (Robbins and Jack, 2006, 80n) "All of these terms denote the capacity to ascribe intentional mental states and to predict and explain behavior on the basis of those ascriptions." (Robbins and Jack, 2006, 62) Similarly, i.e. loosely and approximately but I believe usefully, I would add cognitive empathy (Section 3.2) to the set of terms that denote that capacity. The correspondence between the physical stance and folk physics is rather approximate as well: the stance encompasses scientific as well as folk thinking. However, given present concerns, this, again, is not a problem.

The authors review psychological research on both mindreading and folk

physics and conclude that "Dennett's philosophical distinction between the physical and intentional stances has a lot going for it." (Robbins and Jack, 2006, 64) They find further support for the mindreading/folk physics dichotomy in fMRI studies of the neural regions active during relevant tasks. Additionally, they cite studies of autism as providing evidence for the dissociability of these functions, in that mindreading ability tends to be significantly impaired, while folk physics is unaffected. Research on Williams syndrome seems to suggest dissociation also in the opposite direction. (Robbins and Jack, 2006, 65)

More evidence that "the intentional and the physical stances correspond to very different, even disjoint, modes of construing the world" (Robbins and Jack, 2006, 65) is presented, but perhaps the point has already been made. The authors now introduce their argument for a third mode, turning again to research on a pathological condition for support, this time that of psychopathy, saying that "The lack of concern shown by psychopaths suggests a specific deficit in their empathetic responses to others." (Robbins and Jack, 2006, 67) These authors like others (Section 3.2) find empathy to be difficult to define. For this context, they settle on Frith's concept of "instinctive empathy," as "a basic emotional response that just spills out." (Frith, 2003, 11) Frith contrasts that with what he calls "intentional empathy," which is "affectively neutral and does depend on mindreading: it involves understanding the source of the other persons distress and producing an appropriate behavioral response based on that understanding." (Robbins and Jack, 2006, 67) Robbins and Jack note that Baron-Cohen (2003) "distinguishes between 'affective' and 'cognitive' components of empathy, along similar lines" (Robbins and Jack, 2006, 80n), and it seems quite clear to me that autistics are generally deficient in cognitive, and psychopaths, in affective empathy, though Robbins and Jack go on to offer much more evidence. Yet another term is introduced when they suggest the existence of

> . . . a core component of the normal understanding of what it is to be human—spared in autism but absent in psychopathy, and subserved to some extent by special-purpose neural circuitry—that is relatively independent of mindreading and the intentional

> stance. Tipping our hats to Dennett, we'll call it the "phenomenal stance." (Robbins and Jack, 2006, 69)

Just as regarding something as an intentional system means ascribing intentional states to it, so to regard something as a phenomenal system is to ascribe phenomenal states to it, and this is part of what it is to regard something as "a locus of experience." But it involves more than mere "rote ascription" of such states, "it requires a felt appreciation of their qualitative character. For example, if you don't know what it's like to feel sad, you can't understand what it is [for someone else] to feel sad." (Robbins and Jack, 2006, 70)

Robbins and Jack go on to argue that emotional sensitivity—the sharing of others' pleasure and pain—and moral concern, are also important components of what it is to regard something as a locus of experience, and both involve affective empathy. I find the argument quite compelling as regards affective empathy, and am attracted by its "packaging" as an addition to Dennett's array of stances, but phenomenality for me is problematic—see below and Chapter 4.

Robbins and Jack refer to one of Dennett's early writings on the intentional stance, in which he himself considers a fourth stance:[1] "The personal stance presupposes the intentional stance—and seems, to cursory view at least, to be just the annexation of moral commitment to the intentional." (Dennett, 1981, 240)

They approve of the moral dimension, but stress the differences between the personal and phenomenal stances. One is that, unlike the Dennettian person, the phenomenal system is essentially a subject of conscious experience. Another difference is that the personal stance is inherently intentional, while the phenomenal stance is not, and this, for me, is its major failing. In Section 3.1 we saw that Ward et al. (forthcoming) "ground consciousness in intentionality," in their action-space account, and lose qualia. In my view, with a thorough-going appreciation of intentionality, phenomenality goes the same way (see Chapter 4). This is not to say that I prefer Dennett's per-

[1] The phenomenal stance is a third mode (see above), while the personal stance is the fourth, because Robbins and Jack do not count or discuss the design stance.

sonal stance, though. I believe that the concept of the phenomenal stance is superior in its recognition of the significance of affective empathy. My own stance concept combines that with the intentionality of the personal stance, dropping phenomenality, but before going on to deal properly with that, in the next chapter, I will discuss another paper, on phenomenality, affective empathy and the attribution of sentience/consciousness.

Sytsma and Machery (2009) argue that phenomenality is a preoccuption of philosophers, and not a folk concept. They suggest instead that "For the folk, subjective experience is tightly linked to valence."[2] Sytsma and Machery (2009, 2) (I do not intend to dwell on the former, negative aspect of their thesis, and in fact will deal with the paper relatively briefly, because its function here is merely supplementary to the main arguments against phenomenality and for the significance of affective empathy in the attribution of sentience.) The suggestion of such a link is not original, of course, and they acknowledge Robbins and Jack (2006) (see above, this section), among others, as saying something similar. (Sytsma and Machery, 2009, 7–8)

Sytsma and Machery designed and carried out a number of experiments to test various related hypotheses. Participants, divided into professional philosophers and "ordinary people," were presented with certain scenarios and invited to ascribe mental attributes to the robots and people depicted. (In focusing on the positive aspect of the paper's main thesis, I will ignore the results regarding the philosophers.)

In the first experiment, there appeared a strong tendency to distinguish between seeing red and feeling pain, in that people were willing to allow that a robot might do the former, but not the latter. (Sytsma and Machery, 2009, 12) A second experiment was designed to test the hypothesis that that distinction is due to seeing red being "external" and feeling pain "internal," in which case people should be willing to allow that a robot might be able to smell bananas, but unwilling to consider that it might feel angry. As it turned out, the participants confirmed the hypothesis regarding anger, but were ambivalent about smell. As in the discussion of the first experiment, the authors consider a number of possible objections, but, on

[2]For these authors "mental states have a valence if and only if they have a hedonic value for the subject." (Sytsma and Machery, 2009, 2n)

the assumption that smelling bananas, unlike seeing red, has hedonic value (being pleasant), conclude

> We hypothesize that it is not whether a mental state is the product of the external senses that matters for the folk understanding of subjective experience, but whether they associate that state with some hedonic value for the subject. (Sytsma and Machery, 2009, 23)

The robot might perceive the smell, but is incapable of enjoying it. On the other hand, whereas valence is inherent in the states of pain and anger, it is not so closely tied to smelling bananas, hence the ambivalence. A third experiment was designed to test the hypothesis that there would be less ambivalence in the case of a more "neutral" odor, people being more willing to imagine that a robot might smell it. The third experiment varied odors, using familiar pleasant, familiar unpleasant and unfamiliar ones, predicting that to smell an odor without an associated valence would be more readily attributed to a robot, and the prediction was confirmed by the results. (Sytsma and Machery, 2009, 28)

Though they acknowledge that much work remains to be done, Sytsma and Machery conclude that "for the folk, subjective states seem to be primarily states with a valence." (2009, 31) And to see as significant the valence of states experienced by others—or even, for that matter, to see others' states as having any hedonic value—is, of course, to exercise affective empathy.

The relevance of experimental philosophy in this context is due to the Dennettian position that what require/s explanation is/are the concept/s of consciousness, and though we could argue as to whether the concepts of the philosophers or the folk are the more important, it seems reasonable to deal with both.

3.4 The empathic stance

I was first spurred to think along these lines by a personal observation. Dennett, writing about the intentional stance, suggests that we understand others' actions using rationality, by asking ourselves what a rational agent

would do given the beliefs and desires that we take the other person to have, or deducing beliefs and/or desires given their actions. However, it seems to me that, in many cases, I seek to understand others by asking myself not what a rational agent would do, but what I would do, in their place. This is an example of the empathic stance, and it utilises more of my resources than the intentional one, in particular imagination and subpersonal processing: I imagine myself in that situation and then observe my reactions, which might well be "instinctive" (using that word in a loose sense) or intuitive. It engages my emotions directly, unlike the intentional stance, which "coldly" (Robbins and Jack, 2006, 67) attempts to encapsulate and intellectualise emotion using the concept of explicit desire, while actual, felt emotion is often quite unpredictable. This is subpersonal simulation rather than personal level theory of mind, and involves affective rather than cognitive empathy. (Section 3.2.) Using the intentional stance, the behaviour of creatures that do not think (or certainly not as we do), and therefore are not conscious in Dennett's sense of that word, can only be understood in general, evolutionary terms, so they are seen as mere sensitive mechanisms (Section 1.3.2), but using the empathic stance, they can be seen as sentient individuals that are motivated to seek pleasure and avoid suffering.

The intentional stance, however, involving cognitive empathy as it does, is also an empathic stance of sorts. This terminology has obvious potential for confusion, but I believe the risk is worth taking. I think it reasonable to suggest that, when the word "empathy" is used generally, affective empathy rather than the cognitive variety is what it is most commonly taken to imply. In fact, I would say, the rational nature of operation of the intentional stance makes the use of the word "empathy" in relation to it somewhat counter-intuitive. But this is a purely semantic issue, that I do not believe should cause any significant difficulty.

The empathic stance is very similar to Robbins and Jack's phenomenal stance (Section 3.3), the main (or perhaps only) differences being that the empathic stance is intentional, and the stance taker does not project phenomenality onto the object of the stance, but merely empathises with it. When this experience is rationalised—and in humans the experience always preceeds the rationalisation—some way of designating the type of en-

tity with which empathy is possible is desired, and we call them “sentient” or “conscious”—and a philosopher might mention “phenomenal consciousness,” “phenomenality” and/or “qualia.” (See Section 4.1.) In the context of Dennett’s philosophy, though, we might restrict the term “conscious” to those entities capable of conscious thought, so an appropriate object of the empathic stance is sentient, though they may also be conscious in Dennett’s sense.

Chapter 4

Modeling and Mentality

In the first section of this chapter I consider what the empathic, intentional and design stances have in common, which is second order mental modeling, and on that basis suggest a new "theory of mind" in the traditional philosophical sense of that phrase. In the second section I look at Metzinger's concepts of selfless consciousness in the context of his "self-model theory of subjectivity" (2004), and the third is concerned with self-models more generally.

4.1 Models and minds

In the paper discussed in Section 3.1, Myin and O'Regan suggest that their approach "offers the prospect of 'naturalizing phenomenology'." (2002, 27) Though that particular expression does not appeal to me, I am engaged in a similar exercise, believing that the ideas I present in this dissertation, and particularly in this and the next chapter, go a long way, at least, towards naturalizing sentience. In this section I consider the minimal requirements for the attribution of sentience, or "mindedness" (which I take to be the same). My main conclusions are that intentionality should be considered a feature of models and nothing but models, and that a minimal mind is an embodied system that employs at least two models, one representing beliefs and the other desires, and is affectively motivated to act to reconcile the differences between these models. Some minds, including the human mind,

are also second order modelers, i.e. they model other models (most often and most significantly, other minds). The human mind also models itself, but I address that issue in Sections 4.2 and 4.3. (For me, as for Minsky, below, "modeling" is a general term that includes simulation and carries no implication as to level, i.e. personal or subpersonal.) This can be classed as a "strong representationalist" theory of mind (Lycan, 2006).

A number of writers have worked with the concept of mental models, notably including Johnson-Laird (1983) and Metzinger (2004), but the former focuses tightly on cognition and, while the latter's self-model theory was excluded from proper consideration due to resource constraints, I briefly review aspects of it in the next section. For current purposes, concerning the principles that apply to all modeling, a good starting point is in the work of cognitive scientist Marvin Minsky. Writing in the 1960s, Minsky states that

> If a creature can answer a question about a hypothetical experiment without actually performing it, then it has demonstrated some knowledge about the world. For, his answer to the question must be an encoded description of the behavior (inside the creature) of some sub-machine or "model"...
>
> We use the term "model" in the following sense: To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A.
>
> The model relation is inherently ternary. Any attempt to suppress the role of the intentions of the investigator B leads to circular definitions or to ambiguities about "essential features" and the like. (Minsky, 1968, 425–6)

This is strongly reminiscent of Dennett's intentional stance. Like modeling, intentionality is "inherently ternary," a matter of interpretation: the concept is projected by a third party onto the system of interest. This should not surprise us, however, because modeling and intentionality are very closely related, at least on Minsky's and Dennett's accounts of them. Recall Dennett on the intentional stance:

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure

> out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. (Dennett, 1987b, 17)

This clearly accords with Minsky's concept of a model: to predict an object's behaviour is to "answer a question about a hypothetical experiment without actually performing it." However, this is rather different from a simple physical model, using which, for instance, we predict that a rock that is suspended by hand and then released will fall to the ground. Because they are themselves intentional, the projection of beliefs and desires onto the modeled object means that *it* is being considered a model, so the system that adopts the intentional stance is of a special type: it models models. Dennett calls such a system a "second order intentional system." (1987a, 243)

The concept of the model clearly implies intentionality in Brentano's broad sense of aboutness (Section 1.2.1): that is the relationship between model and object. Models are necessarily intentional, but so are some things that do not seem reasonably to be considered models in themselves, like this text. I would suggest that such things are considered intentional by virtue of the parts they play, or potentially play, in our models, so their intentionality is derivative. And I would go further, to suggest that nothing but a model is intrinsically intentional. There is, however, an important difference between saying that models are intrinsically intentional, and that some systems are intrinsically intentional, which Dennett takes great pains to deny. (1987a) Intentionality is a feature of models by definition—that is the relationship between A* and A that permits the former to answer questions about the latter—but whether a particular system instantiates a model is a matter of interpretation.

(There is an obvious distinction to be drawn between models and entities that employ them; I argue below that a mind necessarily employs at least two models. So we cannot simply translate "intentional system" as "model." Intentional systems derive their intentionality from the models that we interpret them as instantiating.)

So how are we to understand this? Take the example of sight. This minimally involves a light source such as the sun, atmospheric conditions, characteristics of the object concerned, the reflective properties of other objects around it, the various components of the eye, the optic nerve and the brain. The physical information of the eye, the light entering it, and the relevant parts of the nervous system, can carry information about something else (such as an apple) because it has been affected by that thing, and thus has characteristics (such as colour) that correspond to it—there's a causal connection between them. We have evolved and learned to use such connections to obtain information about any visible object.

Dennett talks in terms of the intentional stance, and Minsky about models, but for both, "aboutness" is an inherently ternary relationship: nothing is intrinsically, in and of itself, about anything else, it is only considered to be so by a third party. This arrangement is necessarily uncertain, and there might in fact be no connection whatsoever between the model and what is believed to be its object, if the latter is misidentified. But where there is, in the form of a chain of causation that carries information, that connection even in such a simple case as seeing an apple is indirect, complex and contingent, while the notion of aboutness arches over all such difficulties, suggesting that there is some sort of direct relationship, while there is no such thing. We think of the model–object relationship as continuing, but the real connection, if there is one, lasts only as long as the causal chain that initially carries the information about the object to the model. Of course the more interesting models are dynamic, and can predict changes in the object, but even in successful cases of prediction, the continuing model–object relationship is entirely in the mind of the observer, and the fact that one tracks the other is entirely due to the initial conditions: think of two clocks, both accurate, one of which is set to the same time as the other, and then taken elsewhere. Feedback arrangements, where the model can monitor the object and check prediction outcomes, are much more complex, but the same principles apply. The familiarity of such arrangements, such as my ongoing monitoring of the results of my typing on the computer screen, might lead us to suppose that all intentional relationships are on-going. In the mind of the observer, of course, they are, and that attitude can be extremely useful.

For instance, we might reasonably hope to determine the time indicated on the distant clock by looking at the one nearby, but we should bear in mind that they're not actually connected, either might have lost accuracy, or the distant one might have travelled great distances extremely fast, making it "slow." Such "aboutness" has no observer-independent existence, there are only indirect, complex and contingent flows of information, that the concept of intentionality or modeling greatly simplifies, at best, at the risk of encouraging a spurious degree of certainty. I believe that this risky simplification is the key to understanding intentionality.

Brentano called intentionality "the ineliminable mark of the mental" (1924), but I see modeling as better suited to that role due to its superior explanatory power: modeling has a much more obvious connection with functionality than does intentionality. Intentionality is then seen as merely an aspect of modeling, and the intentional stance as an aspect of second order modeling. As for Brentano's "ineliminability": in principle, the physical stance can be used to predict or explain any action. But to treat the system concerned as a modeler will, in the cases of all but the very simplest of biological entities, be much more practical. (See also below.)

Modeling, though not ineliminable, in my view is the nearest thing to a "mark of the mental" that we will ever find: I believe that a non-modeling mind is genuinely inconceivable, as is a non-mental model that does not depend on a mind to interpret it as such.

A mind is necessarily a modeler, but we can interpret systems as instantiating models without viewing them as minds. Discussions of intentionality tend to focus on belief and desire for a reason: they are, I think, necessary attributes of any mind. Minds do not only represent aspects of their environment, they also act upon it—the concept of a mind that is aware of its surroundings but never takes any action whatsoever seems rather vacuous, or artificial. It is not, I believe, too much of a stretch to view all models as having beliefs: how else could one answer questions about something other than itself? However, minds differ from other conceivable models or modelers in that they also possess desires: minds represent not just how things are, but how they'd like them to be, and seek to minimise discrepancies: to change things to make them more as they think or feel they should be. So

a mind utilises at least two models, one consisting of beliefs and the other, desires, differences between which tend to generate action. "Beliefs," here, is used extremely widely, to include perceptions, and internal states such as pain and hunger. A desire might be as simple as "not hunger" or "not pain." Whether such a simple desire would really constitute another model is not, I would say, a useful question: the concept is not well-defined; there need be no particular way to decide how many models are involved in any given arrangement; if modeling generally is a matter of interpretation, then so, it seems reasonable to suppose, is the number of models employed by a given system. But it certainly makes sense to draw a clear line between beliefs and desires.

So we have two models, the difference between which generates action. We obviously need both a material substrate to instantiate the models and mechanisms to interface with the world: to receive information from and act upon it. In broad terms, we can consider both of these requirements satisfied by embodiment.

For Dennett, modeling first gets off the ground at the Popperian level, with the imagining of alternative potential courses of action and their probable consequences (Section 1.2.2). But intentionality begins at a much lower level, the lowest that he deems worthy of naming, in fact, the Darwinian, which includes the thermostat. I use the concepts of models and modeling in a different way, following Minsky: even a simple mercury thermometer instantiates a model, of current temperature, in that we routinely "ask" the thermometer about the temperature, by looking at it, and it reliably gives us the information we seek. It could be argued that the thermometer does not have a "belief" about the room temperature, it merely indicates its own, which is normally the same. But the point is not how it works, but the way in which we use it, which is to learn the temperature of the room, so, for us, it is a model, albeit of the simplest possible kind (I think). (Perhaps, to be more accurate, we should consider the model to consist not just of the thermometer but also the physical mechanisms due to which its temperature and that of the room generally are linked.) And, being a model, it is an intentional system, even though, unlike the thermostat, it has no desire.

The thermometer, therefore, certainly does not constitute or have a

mind, but does the thermostat? I would suggest that one more element is required for the "natural" attribution of mindedness, and that is affect. Despite some potential affective implication in "desire," it is perfectly clear to anyone of normal mental capacity that the action of a thermostat is entirely mechanical, and the attribution of belief and desire to it is therefore perceived as pedagogical and more metaphorical than literal. But what about cases in which action appears motivated by attraction to pleasure or avoidance of suffering? Perhaps, with some ingenuity, the thermostat could be viewed in that way, but that would be an entirely intellectual exercise, whereas we understand the attraction of cheese for a mouse or the pain of one caught in a trap, in a very different, quite visceral way. So the mouse is considered sentient and the thermostat is not: for a "natural" attribution of sentience, or mindedness, affective empathy is required, in accord with the arguments of Section 3.3.

With mice, of course, (it seems reasonable to say) we share not only the capacities for pleasure and suffering generally, but the specific experiences of hunger and pain. That is not the case when we consider the simplest of organisms, such as bacteria and plants, but such primitive reactions as tropism and taxis can be viewed as the organism moving towards what it likes, and away from what it dislikes, without distorting our understanding of them, even if more detailed investigation inevitably requires descent to the physical stance.

So we are now saying that, to be considered sentient, a system should have beliefs and desires, and the motivation to act, to realise the desires, should be seen as affective rather than merely mechanical. We can go along with Dennett, take the design stance, and view the attraction of cheese for mice as rational, because that is how evolution arranged matters, but the mentality of the individual mouse is obviously not rational. Given what he has said about sentience (Section 1.3.2), for Dennett the mouse is merely a sensitive mechanism, but I view it as a creature motivated to seek pleasure and avoid suffering, which I think is more ethical and no less realistic, as long as implausible and unnecessary considerations are not invoked to explain it. We can, of course, easily imagine a sentient creature in a perceptual state of neutral hedonic value, but if that state was seen as permanent, so

that no degree of pleasure or suffering could ever be experienced, then I think the creature would be reduced in our minds to nothing but a sensitive mechanism.

(Similar considerations can be applied to non-conscious mental states: if *all* of a system's states are non-conscious, i.e. it is not a sentient or conscious system, then none of its states would be considered mental: the mentality of the non-conscious states depends on other states of the same system being conscious or sentient.)

Dennett's design stance (Section 1.2.1) can, I think, be analysed in terms of modeling. The designed object is intentional in the sense that it refers to the task or situation for which it was designed, and so (see above) it plays a part in at least one person's mental model(s)—or those of at least two people, if the designer is counted, and the potential user is a different person. (Evolution, of course, as an impersonal designer, is an atypical example.) Empathy is involved where the potential user attempts to use the thing "properly," i.e. in the way envisaged by the designer, and, of course, this is cognitive rather than affective empathy. I see the design stance as an application of second order modeling, in which the user models the designer—albeit in quite a minimal way, in most cases.

The attribution of sentience is, at the most basic level, the modeling of a first order modeler by a second order modeler, in which the latter views the former as having beliefs and desires and being affectively motivated to reconcile these (or at least such attribution signifies that the system of interest is considered suitable for such modeling). This, of course, is where affective empathy comes into the picture. To view mouse behaviour as evolutionarily rational, using concepts such as hunger and pain rather than descending to the physical stance, is to employ cognitive rather than affective empathy.

4.2 Metzinger's concepts of selfless consciousness

In his book Being No One, Thomas Metzinger repeatedly emphasises his claim, implied in the title, that "no such things as selves exist in the world: Nobody ever *had* or *was* a self." (2004, 1, Metzinger's emphasis) However, an important part of his "self-model theory of subjectivity" (the book's

subtitle) is what he calls the "phenomenal self-model" (PSM), which might loosely be characterised as the appearance of self, or our experience of selfhood: it is the main reason that we generally believe that we do, indeed, have or are selves.

This book is a very substantial work, and the self-model theory (SMT) is complex and sophisticated, and the limited resources available for this research necessitate that we attend only to the most directly relevant aspects, from which the concept of "selfless consciousness" (2004, 563–567) stands out. Of course, what this means, given Metzinger's main thesis, is consciousness that lacks the PSM. He describes two varieties of selfless consciousness that he considers theoretically possible, of which one maps very neatly onto the concept of sentience. The other certainly looks as if it should relate to Blackmore's concept of "selfless" meditation (Section 2.2)—Metzinger actually identifies it with the Buddhist concept of enlightenment (2004, 566)—but given the scale of the SMT, an attempt to fully flesh-out that relationship lies beyond the scope of this dissertation. I do, however, attempt to outline it below.

Before dealing with these concepts of selfless consciousness, though, I should perhaps say something about Metzinger's extensive use of phenomenality in his theory. This is basically a functionalist approach, and it therefore falls at the same hurdle as any other functionalist approach, as described in Section 3.1.In principle, it should be possible to reformat the SMT in third-person terms, but it would then look very different, and there's at least a theoretical possibility that it might be reduced to insignificance, because what it seeks to explain—selfhood and subjectivity—are themselves subjective phenomena. That is not to say that the SMT might not be a highly fruitful theory—I do not share Dennett's commitment to "the objective, materialistic, third-person world of the physical sciences." (1987a, 5) But the one thing that it necessarily cannot do, as a functionalist theory, is to naturalise phenomenality.

The SMT is constraint-based, and according to it, normal phenomenal consciousness in humans satisfies ten constraints, of which only a subset directly concerns us here. The more simple of the two varieties of selfless consciousness, Metzinger himself clearly views as sentience, suggesting that

"many simple organisms on our planet belong to this phenomenal system class." (2004, 563) Indeed, that could almost function as a definition of sentience. Such a system will satisfy constraints 2, 3 and 7 but will not possess a centered model of reality. "Centered" is important, there, because, according to Metzinger, such systems will have a world-model, though no self-model.

So, what is the subset of constraints that sentience will satisfy? In addition to those constraints that are directly implicated I will briefly mention constraint 1. This is functional rather than phenomenal, being the "global availability" of information in the system "for deliberately guided attention, cognitive reference and control of action." (2004, 118) As Metzinger acknowledges, this has previously been suggested as a defining characteristic of consciousness by Baars (1988; 1997) and Chalmers (1997). Perhaps Metzinger omits global availability from the requirements for sentience because of the awkwardness of applying such concepts as deliberate guiding, control and action where there is not even the appearance of a self to guide, control or act. This does seem awkward for him, but I will not pursue it here.

Constraint 2 is "activation within a window of presence." (2004, 126) This is a temporal feature, and Metzinger classifies it as primarily phenomenological rather than functional: it means that consciousness is very closely tied to the present: whatever is experienced, is experienced *now*. "Only persons possessing a subjective Now are *present* beings, for themselves and for others." (2004, 126) One might suppose that, in a state of selfless consciousness, one is a present being only for others, but if that state might be attained temporarily, for instance in meditation, the situation is more complex—but I am getting ahead of myself—see below. Due to short-term memory, the experience of the present is not instantaneous, but extends over a period of time, and this is what is meant by "a window of presence." (2004, 129)

Constraint 3 is "integration into a coherent global state... Individual conscious states, in standard situations, are always part of a conscious world-model... Consciousness is... 'being-in-the-world'." (2004, 131) When he wrote that, Metzinger was probably thinking of normal consciousness; selfless consciousness might be better characterised as merely "in-the-

world." But the integrated world-model that is common to both normal and selfless consciousness is the main point here.

Constraint 7 is transparency. This applies only to a subset of phenomenal representations, but for Metzinger it is vital: "From a systematic point of view, and in particular for the main argument here, the transparency constraint is of the highest relevance." (2004, 163) He notes that a concept of phenomenal transparency is in quite common use among contemporary philosophers, but omits any mention of Clark (2000) (Section 3.1), in which it plays an important part. He differentiates his concept from the common one: that property is typically considered to apply to all phenomenal states, while for Metzinger, "For any phenomenal state, the degree of phenomenal transparency is inversely proportional to the introspective degree of attentional availability of earlier processing stages." (2004, 165) Where the common concept implies availability only of contents, not their vehicles, Metzinger makes that distinction a matter of degree, and variable between types of phenomenal state.

So, according to Metzinger, sentience is active within a window of presence, integrated into a coherent global state, and transparent.

In the other, more complex state of selfless consciousness, achievable (in principle, at least) by humans, there is a phenomenal self-model (PSM), but it is fully opaque. What this means is that it is apprehended *as* a model, a representational construct, and therefore is viewed not as "the self," but as a model of the system. Metzinger makes a "phenomenological analogy," that of the lucid dream. (2004, 565–6) The lucid dreamer recognises the dream as such: a representational construct. Metzinger asks us to imagine a lucid dream in which *the dreamer* is recognised as a dream character, a representational fiction. The second state of selfless consciousness is similar, at least as regards the concept of the self, but occurs in waking life.

Tempting as it is to further investigate these fascinating ideas, they would take us too far from our present concerns. Certainly, Blackmore's concept of meditation shares apparent selflessness with these states described by Metzinger—and, in other writings, Blackmore analyses the Buddhist concept of enlightenment in similar terms (1999)—but meditation as the cessation of memetic activity, which is what we are working with here, despite

being, according to Blackmore, a selfless state of consciousness (Blackmore, 2003, 25), cannot easily be related to Metzinger's opaque PSM.

4.3 Attribution to self and others

As implied in the previous section, in normal human consciousness Metzinger's PSM is transparent, and so we seem to perceive, and form a concept of, the self. But we obviously have such a concept, whether it develops in the way that Metzinger suggests or not, and I would go further, in stating that it should be fairly obvious that we each have a self-model, however closely or otherwise it resembles the PSM. (See the paragraphs on Minsky, below.)

I believe that the attribution of sentience/consciousness amounts to projection of the self-model, i.e. identification with the attributee, or empathy. The philosophical zombie has previously been described as an imaginary person with whom we imagine having absolutely no empathy (Section 3.1). It could equally be viewed as an imaginary body that is animated despite lacking an inhabiting self. This, of course, accords with the link between the self and consciousness in Dennett's thinking (Section 1.2.3).

However, it might seem wrong to project self-hood in the case of sentience, or in Blackmore's concept of selfless meditation, because here there is no self-model. That would be a mistake, due to identifying the self with the self-model—recall that, for Metzinger (as well as for myself, ironically), there is no self in any case, and when the PSM is opaque, i.e. is seen for what it is, it becomes a system model rather than a self-model, so the sentient entity "merely" lacks a system model. But with affective empathy, we recognise patterns of behaviour that we associate with positive and negative hedonic states, and we identify with the creature that is seen as experiencing them. (And, when a particular creature or type of creature is considered sentient, we view them also as experiencing states of neutral hedonic value.)

Minsky also considered self-models:

> When a man M answers questions about the world, then (taking on ourselves the role of B [the observer]) we attribute this

ability to some internal mechanism W* inside M. . .

> But what about broader questions about the nature of the world? These have to be treated (by M) not as questions to be answered by W*, but as questions to be answered by making general statements about W*. If W* contains a model M* of M then M* can contain a model W** of W*; and, going one step further, W** may contain a model M** of M*. Indeed, this must be the case if M is to answer general questions about himself. Ordinary questions about himself, e.g., how tall he is, are answered by M*, but very broad questions about his nature, e.g., what kind of a thing he is, etc., are answered, if at all, by descriptive statements made by M** about M*. (1968, 426–7)

It might be helpful to restate (my understanding of) these ideas in slightly different form.

Assuming that all healthy adults can answer questions about the world, we all have world models. But broad questions about the nature of the world are answered not by the model but by statements about the model (implying a second order model).

My world model hypothetically contains a model of myself, that model might contain a model of my world model, and that in turn might contain a model of my model of myself. Necessarily so, in fact, if I am to answer general questions about myself, as opposed to specific ones, which utilise only the first level self-model.

Minsky continues:

> The reader may be anxious, at this point, for more details about the relation between W* and W**. How can he tell, for example, when a question is of the kind that requires reference to W** rather than to W*. Is W** a part of W? (Certainly W*, like everything else, is part of W.) Unfortunately, I cannot supply these details yet, and I expect serious problems in eventually clarifying them. We must envision W** as including an interpretative mechanism that can make reference to W*, using it as a sort of computer-program subroutine, to a certain depth

> of recursion. In this sense W** must contain W*, but in another, more straightforward, sense W* can contain W**. This suggests first that the notion "contained in" is not sufficiently sophisticated to describe the kinds of relations between parts of programlike processes and second that the intuitive notion of "model" used herein is likewise too unsophisticated to support developing the theory in technical detail. It is clear that in this area one cannot describe intermodel relationships in terms of models as simple physical substructures. An adequate analysis will need much more advanced ideas about symbolic representation of information-processing structures. (1968, 427)

It is tempting to suppose that part, at least, of Minsky's difficulty as expressed in that paragraph is due to his total reliance on information processing concepts, unlike Metzinger's self-model theory, in which embodiment plays an important role. However, I think that Minksy makes an important contribution in flagging up the significance of higher order models. My own position is quite closely related to higher order theories of consciousness—see, for instance, Carruthers (2001)—in that I view the concepts of sentience and consciousness as being clearly located in a higher order self-model.

Chapter 5

Information and conclusions

This, the final chapter, is rather more speculative than its predecessors, and I mention a number of ideas that, to be treated "properly," would require many more words than I can devote to them here, but they are all closely related to the main themes of this dissertation, in particular the naturalisation of phenomenal consciousness.

In Chapter 4, the concept of mind was analysed in terms of models, so that a system that can reasonably be considered to "possess" a mind, or be sentient (at least, if not also conscious), is necessarily a modeler. The next logical analytical step, I believe, given the nature of models, is to focus on information.

The function of the first section below is to explain "physical" and "intentional" information. This might be considered a dual aspect theory of information, but it differs from that of Chalmers (1996), for instance, because the aspects are distinguished not metaphysically, but psychologically.

5.1 Two aspects of information

> The word "information" has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of

> information would satisfactorily account for the numerous possible applications of this general field. (Shannon et al., 1993)

I quote Shannon not to justify my use of two concepts of information, but to help explain why I do not argue that my concepts are better than others, such as that of Dretske (see, for instance, his (1981; 2000)). For Dretske, information is akin to meaning, but whereas a meaningful statement might or might not be true, information, to be informative, must be true. While I agree that is a common and, indeed, useful concept, I believe that my use of the term is better suited to the sort of theoretical analysis that I attempt below. The more important of the many uses of "information," including physical information, are surveyed in Floridi (2004).

At a physics conference in Dublin in July 2004, Stephen Hawking admitted he'd made a mistake. (Preskill, 2004) He had bet John Preskill, another physicist, that black holes destroy information, but Hawking changed his mind, conceded publicly at the conference, and presented Preskill with his prize: an encyclopædia (to represent enduring information). Hawking had come to believe that information is preserved inside black holes, and eventually escapes them. A black hole, he now agreed, behaves like a computer: input information is transformed into output information in a systematic way that could, in principle, be deduced by comparing input and output. Anything that falls into a black hole is considered input information. The output is in the form of "Hawking radiation," an earlier theoretical discovery of his. If it were possible to monitor all that falls into a black hole, and all of the Hawking radiation that emerges from it, it would be possible in principle to discover all that goes on within it.

According to physicist Roy Frieden (1998), the laws of physics are generated by the attempt to minimize the difference between an entity or system's own physical information, and the information that physicists can obtain about it. That particular suggestion remains controversial, but the concept of physical information is now very well established (see, for instance, Leff et al. (1992)). It developed from the concept of information in communication theory (Shannon, 1948), and like that, has no semantic aspect or component. Physical information is basically material form, which, using

mathematical techniques derived from communication theory, can be quantified. It might be thought of as shape, but the concept is actually much more general, encompassing all physical qualities. So any physical entity can be considered to embody its own description, if its form (in that general sense) is treated as physical information.

To experiment upon a physical object is to seek to extract its physical information from it, but there are two complicating factors that need to be considered: once obtained, information *about* the object is not physical information, despite its referent being entirely physical, and the fact that, in the ideal case, it perfectly duplicates the physical information. The other factor is that the information obtained directly from the experiment actually concerns not the object alone, but the interaction between that and the experimental apparatus, whatever that might be. To obtain information *about* the object, further processing is required.

This brings us to an important principle: every physical thing encodes the outcomes of all of its potential interactions. This is most easily envisaged using a simple case, where there are just two things, such as two asteroids on a collision path: each of them can be considered to encode the aftermath of the collision, where the other is the decoding key.

Now, what is encoded anywhere is necessarily (but loosely speaking) "a matter of opinion." What I mean is that it depends on the decoding key. Of course, in the case of human communications, a particular message is intended to be conveyed, and that is the criterion of correct decoding. However, the concepts of en- and decoding are generalised here, being applied to natural phenomena, where that criterion obviously does not apply. (It is a trivial consequence of considering material form as information that later states of affairs can be considered encoded in earlier ones, but I have not found any literature in which the same sort of weight is placed upon this concept as here.) In the case of the asteroids, either one can be considered to encode the outcome of the collision, simply by designating the other as the key. This is why every physical entity can be considered to encode the outcomes of all of its potential interactions. The only constraint is context (in this case I do not distinguish key from context—see below). The cellular context is what limits the outcome of DNA decoding to (generally)

biologically valid configurations, and the outcomes of memetic en/decoding processes are similarly constrained.

Turning again to our asteroids, in reality there will always be influences also from other objects, such as gravitation. Ultimately, every physical thing is continuously interacting with the rest of the universe. However, some interactions are obviously more significant than others. The distinctions between decoding key, context and background are entirely relative: the most important interacting entity will usually be considered the key, the less important but non-negligible entities constitute the context, and those that can safely be ignored are relegated to the background. But in some cases we need to allow for many influences, and there we can consider the context to be the decoding key. This is how I tend to think of the en/decoding of memes, where the context includes both internal and external factors.

Genes are items of physical information that are encoded in strands of DNA, to be decoded by the cellular machinery. (Dennett, 1995a) Memes (Section 2.1), similarly, are items of physical information that are encoded in brains and in behaviour (and, via behaviour, in symbols, artefacts and other behavioural traces). A brain that contains a particular meme, suitably stimulated, will produce the relevant pattern of behaviour that, in turn, might be observed and consequently encoded within another brain.

> It is tempting to suppose that some concept of *information* could serve eventually to unify mind, matter, and meaning in a single theory. (Dennett and Haugeland, 1987, emphasis in the original)[1]

As stated above, information obtained by experimenting upon a physical thing is not in itself physical information (which is not to say that it is not physical—see below). This is what I call intentional information: information that is about something, whether that thing is real or not.

For me, the concept of intentional information, inspired by the passage quoted above, has two major virtues: it is a general term, encompassing all meaning, reference, significance, etc.; and the relationship between it

[1] There is a version of this text on the web from which the relevant paragraph has been cut, and it is also missing from later editions of the book.

and physical information—and therefore, ultimately, between meaning, etc. and matter—can be stated quite simply: intentional information is always encoded in physical information. Genes and memes are similarly encoded, though, so what distinguishes intentional information? Intentionality has, of course, already been discussed, mainly in Section 4.1. So intentional information is that which plays a part in a model.

But the crucial point is this: it is, I believe, the encoding of intentional information in physical information, and the consequent context-dependent nature of it, that gives subjectivity its defining characteristics: uncertainty, relativity to point of view, and so on. In particular, the context-dependent nature of en/decoding can be considered to explain the dependency of intentionality and modeling on interpretation (Section 4.1). However, which of these aspects is more fundamental is perhaps a matter of opinion.

What is the ontological status of intentional information?

I hope that it is safe to assume that the reality of physical information or material form is uncontentious. In Dennett's Real Patterns (1991b), he argues, roughly speaking, that a pattern is real *iff* it is compressible (following Chaitin (May, 1975)). Now, this sort of information is neither physical nor intentional, but it shares features with each of these. Like intentional information, it is always encoded in physical information, but like the latter it does not refer to or mean anything[2]—it could be viewed as "pure pattern," like the concept of information in communication theory. So I take from Real Patterns that patterns encoded in physical information can be considered real. However, I believe this position also to be consistent with Ross's "rainforest realism" (Ross, 2000), in which he denies the abstracta/illata distinction that Dennett relies upon.

Intentionality, though, is a matter of interpretation (Section 4.1 and above). That necessarily applies to intentional information *as such*, but not

[2]To view physical information as refering to the object that instantiates it is a mistake, because it *is* the form of that object. I have argued elsewhere (unpublished) that (yet) another stance could be added to Dennett's array, that I call the formal stance, which we adopt whenever we focus on form rather than substance, so that an object's physical information simply is the object, viewed from the formal stance. (The stance that complements the formal one, the "substantial stance," is the physical stance.)

to the patterns concerned. This becomes rather obvious when an appropriate example is chosen: the patterns of paint on a canvas are perfectly real, while the scene depicted might be entirely imaginary. Like the paint, and also memes and genes, the neural and other patterns that serve as the vehicles for intentional information are actually physical, however complex, distributed or otherwise difficult to discern they might be. Ultimately, all information is physical, because intentional information is, strictly, an aspect of the use of physical information, rather than any kind of entity in itself. (The concept of intentional information as the use of physical information was inspired by the later position of Wittgenstein on meaning (1972), of which I think it reasonable to view this as a generalisation.)

That point is (for me at least) a little difficult to grasp, but I think this consideration helps: due to our intimate familiarity with it, intentional information for us is generally quite explicit. However, in "the objective, materialistic, third-person world of the physical sciences" (Dennett, 1987a, 5), intentional information is *always* encoded, or implicit, in physical information, and never appears in clear form. (From the physical stance, intentional information simply does not exist, while from the intentional stance, of course, it does.) Phenomenal consciousness can be considered a stream of intentional information, the experience of which is what it is like to be a certain sort of information processor, one that is a user of information, the minimal example of which is the sentient modeler described in Section 4.1.

In that section, intentionality was illustrated using the example of a person seeing an apple. It might be beneficial here to express that in terms of information. When a beam of light entering your eye carries information about an apple, that information is encoded. The encoding takes place when the light encounters the surface of the apple and is filtered by the structures it finds there as it is reflected, so the balance of the mixture of wavelengths within it is changed. The decoding takes place within the eye, the optic nerve and the brain, as that particular mixture of wavelengths is interpreted to be the colour of the apple. Only the light's own physical information enters the eye, but that can be processed to yield information about the apple, due to the chain of causation that connects these. The physical information of the light is the carrier, the brain etc. is the decoding mechanism, and the

apple's colour is the coded message. Without the intentional stance, there is just the physical information of the brain's structure (however complex), but if we take that stance, some of that physical information can be taken to encode intentional information about things outside the brain. Intentionality uses physical information while obscuring the details of that use.

5.2 Conclusions

This research project was originally stimulated by Blackmore's claim that Dennett's theory of consciousness fails to explain her experience and understanding of meditation, as meme-free. (Blackmore, 2003) This state of, roughly, unthinking awareness reminded me of Clark's suggestion that Dennett's concept of an experiencer seems unduly "thick" (Clark, 2002), and also of Dennett's failure, in my view, to adequately explain sentience.

I therefore set out to find an explanation of sentience that is consistent with the main body of Dennett's philosophy. Robbins and Jack's concept of the "phenomenal stance" (2006), based on affective empathy, came quite close, but I was troubled by the concept of phenomenality, believing it to be too subjective to sit comfortably within Dennett's "objective, materialistic, third-person world of the physical sciences" (1987a, 5), and also a philosophical "artefact," absent from folk philosophy.

By taking a "strong representationalist" standpoint, based on intentionality and modeling, I found I could eliminate phenomenality from the phenomenal stance, leaving us with what I call "the empathic stance." This is, like Dennett's intentional stance, a matter of interpretation, but I believe that it explains the concept of sentience: we experience affective empathy for the apparent ability of "lower" organisms to experience pleasure and suffering, and then rationalise our experience by saying that such organisms are sentient. This answers the concerns of Clark and myself. As for Blackmore, the state of a human who is unthinking, but remains capable of hedonic experience, is conceptually very similar to that of a creature that is incapable of thought (at least as we generally think of it), so we might view a meditator (or at least a person in a state that corresponds to Blackmore's concept of meditation) as sentient.

Finally, using the concepts of material form as physical information (from physics) and meaning, representation, etc. as intentional information (my own concept), I argued that the latter is very usefully viewed as always encoded in the former.

Bibliography

G.W. Allport. *Personality: a Psychological Interpretation*. Henry Holt, New York, 1937. [36]

Robert Aunger, editor. *Darwinizing Culture: The Status of Memetics as a Science*. Oxford University Press, 2001. [19]

B.J. Baars. *A cognitive theory of consciousness*. Cambridge Univ Pr, 1988. [56]

B.J. Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA, 1997. [56]

S. Baron-Cohen. *The Essential Difference: Male and Female Brains and the Truth about Autism*. Basic Books, 2003. [41]

S. Baron-Cohen, A.M. Leslie, U. Frith, et al. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985. [33]

C.D. Batson. These things called empathy: Eight related but distinct phenomena. In Decety and Ickes (2009). [37, 38]

S. Blackmore. Memes and the malign user illusion. In *Association for the Scientific Study of Consciousness Conference Unity and Dissociation, Brussels*, 2000. [19, 23]

S. Blackmore. Consciousness in meme machines. *Journal of Consciousness Studies, 10*, 4(5):19–30, 2003. [2, 19, 22, 23, 24, 58, 67]

S.J. Blackmore. *The meme machine*. Oxford University Press, USA, 1999. [22, 24, 57]

N. Block. On a confusion about a function of consciousness. In Block et al. (1997). [25]

N. Block, O. Flanagan, and G. Guzeldere, editors. *The Nature of Consciousness*. M.I.T. Press, 1997. [70]

M.A. Boden. *Mind as machine: a history of cognitive science*. Oxford University Press, USA, 2006. [34]

F. Brentano. *Psychologie vom empirischen Standpunkt (Bd. 1) [Psychology from an empirical perspective]*. Meiner, Leipzig, 1924. Original work published in 1874. [2, 7, 51]

Peter Carruthers. Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy*, 2001. URL `http://plato.stanford.edu/entries/consciousness-higher/`. Retrieved 29 November 2009. [60]

Gregory Chaitin. Randomness and mathematical proof. *Scientific American*, 232:47–52, May, 1975. [65]

D. Chalmers. The representational character of experience. In Brian Leiter, editor, *The future for philosophy*, pages 153–81. Oxford University Press, 2004. [27]

David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–19, 1995. [25]

D.J. Chalmers. *The conscious mind*. Oxford University Press, 1996. [40, 61]

D.J. Chalmers. Availability: The cognitive basis of experience. *Behavioral and Brain Sciences*, 20(01):148–149, 1997. [56]

P.S. Churchland. *Neurophilosophy: Toward a unified science of the mind-brain*. The MIT Press, 1989. [3]

Andy Clark. A case where access implies qualia. *Analysis*, 60(1):30–38, 2000. [2, 25, 26, 27, 28, 30, 57]

Andy Clark. That special something: Dennett on the making of minds and selves. In A. Brook and D. Ross, editors, *Daniel Dennett*, pages 187–205. Cambridge Univ Pr, 2002. [2, 14, 15, 25, 67]

Christian Coseru. Mind in Indian Buddhist Philosophy. *Stanford Encyclopedia of Philosophy*, 2010. URL `http://plato.stanford.edu/entries/mind-indian-buddhism/`. Retrieved 14 December 2010. [4]

M. Davies and T. Stone. Mental simulation, tacit theory, and the threat of collapse. *Philosophical Topics*, 29(1&2):127–73, 2001. [33, 35]

R. Dawkins. Viruses of the mind. In Bo Dahlbom, editor, *Dennett and his Critics: Demystifying Mind*. Blackwell: Oxford, 1993. [19]

R. Dawkins. *The extended phenotype: The long reach of the gene*. Oxford University Press, USA, 1999. [19]

Richard Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1976. [18, 19]

Frederique de Vignemont and Tania Singer. The empathic brain: how, when and why? *TRENDS in Cognitive Sciences*, 10(10), 2006. [39]

Jean Decety and William Ickes, editors. *The Social Neuroscience of Empathy*. M.I.T. Press, Cambridge, Massachusetts, 2009. [69, 75, 76, 77]

Daniel C. Dennett. Intentional systems. *The Journal of Philosophy*, pages 87–106, 1971. [8, 9]

Daniel C. Dennett. Beliefs about beliefs. *Behavioural and Brain Sciences*, 4: 568–70, 1978. [32]

Daniel C. Dennett. Mechanism and responsibility. In Daniel C. Dennett, editor, *Brainstorms: Philosophical Essays on Mind and Psychology*, pages 233–255. MIT Press, 1981. [42]

Daniel C. Dennett, editor. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts, 1987a. [1, 4, 49, 55, 66, 67]

Daniel C. Dennett. True believers. In Daniel C. Dennett, editor, *The intentional stance*. MIT Press, 1987b. First published in 1981. [8, 10, 49]

Daniel C. Dennett. *Consciousness Explained*. Allen Lane, London, 1991a. [4, 6, 19, 22, 23, 24, 40]

Daniel C. Dennett. Real patterns. *Journal of Philosophy*, 87:27–51, 1991b. [13, 65]

Daniel C. Dennett. The self as a center of narrative gravity. In F. Kessel, P. Cole, and D. Johnson, editors, *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum, 1992. URL `http://ase.tufts.edu/cogstud/papers/selfctr.htm`. Retrieved 28 February 2010. [11, 12, 13]

Daniel C. Dennett. *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster, 1995a. [21, 64]

Daniel C. Dennett. *Kinds of minds: Toward an understanding of consciousness*. Basic Books, 1997. [10, 15, 16]

Daniel C. Dennett and John Haugeland. Intentionality. In Gregory (1987). [64]

D.C. Dennett. *Brainstorms: Philosophical essays on mind and psychology*. The MIT Press, 1979. [3]

D.C. Dennett. The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, 2:322–6, 1995b. [27, 31]

F.I. Dretske. *Perception, knowledge, and belief: selected essays*. Cambridge Univ Pr, 2000. [62]

Fred Dretske. *Knowledge and the Flow of Information*. The MIT Press, 1981. [62]

N. Eisenberg and J. Strayer, editors. *Empathy and its Development*. Cambridge University Press, Cambridge, 1990. [77]

G. Evans. Molyneux's question. In A. Phillips, editor, *Gareth Evans: Collected Papers*. Clarendon Press, 1985. [30]

Luciano Floridi, editor. *The Blackwell Guide to the Philosophy of Computing and Information*. Blackwell, 2004. [62]

B. Roy Frieden. *Physics from Fisher Information*. Cambridge University Press, Cambridge, 1998. [62]

U. Frith. *Autism: Explaining the enigma*. Wiley-Blackwell, 2003. [41]

V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998. [35]

V. Gazzola and C. Keysers. The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex*, 19(6):1239–1255, 2009. [35]

A. Goldman. Interpretation psychologized. *Mind and Language*, 4(3):161–185, 1989. [33]

A. Gopnik and A.N. Meltzoff. *Words, Thoughts, and Theories*. M.I.T. Press, 1997. [33]

R. Gordon. Folk psychology as simulation. *Mind and language*, 1(2):158–171, 1986. [34]

R.M. Gordon. Folk psychology as mental simulation. *Stanford Encyclopedia of Philosophy*, 2009. URL `http://plato.stanford.edu/entries/folkpsych-simulation/`. Retrieved 8 December 2009. [34, 38]

Richard L. Gregory, editor. *The Oxford Companion to the Mind*. Oxford University Press, Oxford, 1987. [72]

R. Grush. Skill and spatial content. *Electronic Journal of Analytic Philosophy*, 6(6), 1998. [30]

J. Heal. Replication and functionalism. In J. Butterfield, editor, *Language, mind and logic*, pages 135–150. Cambridge University Press, 1986. [34]

J. Heal. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language*, 13(4):477–498, 1998. [34, 35]

F. Jackson. Mind and illusion. In A. O'Hear, editor, *Minds and persons: Royal Institute of Philosophy supplement 53*. Cambridge University Press, 2003. [27]

J. Jaynes. *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin Company, 1976. [11, 12]

Philip N. Johnson-Laird. *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*. Harvard University Press, 1983. [48]

Robert Kirk. Zombies. *Stanford Encyclopedia of Philosophy*, 2006. URL `http://plato.stanford.edu/entries/zombies/`. Retrieved 28 November 2010. [27]

H.S. Leff, A.F. Rex, and D.L. Hogenboom. Maxwell's demon: entropy, information, computing. *American Journal of Physics*, 60:282–283, 1992. ISSN 0002-9505. [62]

J. Levine. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(9):354–61, 1983. [40]

T. Lipps. Einfühlung, inner nachahmung, und organ-empfindungen. *Archiv für die gesamte Psychologie*, 2:185–204, 1903. [36]

William Lycan. Representational theories of consciousness. *Stanford Encyclopedia of Philosophy*, 2006. URL `http://plato.stanford.edu/entries/consciousness-representational/`. Retrieved 17 December 2010. [48]

Joan McCarthy. *Dennett and Ricoeur on the Narrative Self*. Prometheus Books, 2007. [13]

T. Metzinger. *Being no one: The self-model theory of subjectivity*. The MIT Press, 2004. [2, 47, 48, 54, 55, 56, 57]

M.L. Minsky. Matter, mind and models. In M.L. Minsky, editor, *Semantic information processing*. The MIT Press, 1968. URL `http://web.media.mit.edu/~minsky/papers/MatterMindModels.html`. First published in 1965, web version retrieved 1 May 2010. [48, 59, 60]

J.P. Mitchell. The false dichotomy between simulation and theory-theory: the argument's error. *Trends in Cognitive Sciences*, 9(8):363–364, 2005. [35, 38]

Erik Myin and J. Kevin O'Regan. Perceptual consciousness, access to modality and skill theories. *Journal of Consciousness Studies*, 9(1):27–45, 2002. [26, 28, 29, 30, 47]

Steven Nadler. Baruch Spinoza. *Stanford Encyclopedia of Philosophy*, 2007. URL `http://plato.stanford.edu/entries/spinoza/`. Retrieved 14 December 2010. [3]

Thomas Nagel. *Mortal Questions*. Cambridge University Press, Cambridge, 1979a. [75]

Thomas Nagel. Subjective and objective. In *Mortal Questions* Nagel (1979a). [3]

Thomas Nagel. What is it like to be a bat? In *Mortal Questions* Nagel (1979a). First published in 1974. [1, 3, 4, 29]

Thomas Nagel. *The View From Nowhere*. Oxford University Press, New York, 1986. [1]

S. Nichols and S. Stich. Second thoughts on simulation. In T. Stone and M. Davies, editors, *Mental Simulation: Evaluations and Applications*, pages 87–108. Blackwell, 1995. [35]

S. Nichols and S.P. Stich. *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford University Press, USA, 2003. [38]

J. Perner and A. Kuhlberger. Mental simulation: Royal road to other minds. In B.F. Malle and S.D. Hodges, editors, *Other minds: How humans bridge the divide between self and others*, pages 174–189. The Guilford Press, 2005. [38]

J.H. Pfeifer and M. Dapretto. "Mirror, mirror, in my mind:" empathy, interpersonal competence, and the mirror neuron system. In Decety and Ickes (2009). [39]

D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain sciences*, 1(4):515–526, 1978. [32]

John Preskill. On Hawkings Concession, 2004. URL `http://www.theory.caltech.edu/~preskill/jp_24jul04.html`. Retrieved 12 February 2007. [62]

Philip Robbins and Anthony I. Jack. The phenomenal stance. *Philosophical studies*, 127(1):59–85, 2006. [2, 40, 41, 42, 43, 45, 67]

D. Ross. Rainforest realism: A Dennettian theory of existence. In Ross et al. (2000), pages 147–68. ISBN 026268117X. [65]

D. Ross, A. Brook, and D. Thompson, editors. *Dennett's philosophy: a comprehensive assessment*. The MIT Press, 2000. ISBN 026268117X. [76]

Gilbert Ryle. *The Concept of Mind*. University of Chicago Press, 1949. [5]

R. Saxe. Against simulation: the argument from error. *Trends in Cognitive Sciences*, 9(4):174–179, 2005a. [35, 36]

R. Saxe. Hybrid vigour: Reply to Mitchell. *Trends in Cognitive Sciences*, 9 (8):364, 2005b. [38]

Simone G. Shamay-Tsoori. Empathic processing: its cognitive and affective dimensions and neuroanatomical basis. In Decety and Ickes (2009). [38, 39]

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948. [62]

C.E. Shannon, N.J.A. Sloane, and A.D. Wyner. *Claude Elwood Shannon: Collected Papers*. IEEE Press, 1993. [62]

J. Sytsma and E. Machery. Two conceptions of subjective experience. *Philosophical Studies*, pages 1–29, 2009. [43, 44]

Michael Tye. Qualia. *Stanford Encyclopedia of Philosophy*, 2007. URL `http://plato.stanford.edu/entries/qualia/`. Retrieved 30 November 2010. [28]

D. Ward, T. Roberts, and A. Clark. Knowing what we can do: actions, intentions, and the construction of phenomenal experience. *Synthese*, forthcoming. [27, 28, 29, 31, 42]

J.C. Watson and L.S. Greenberg. Empathic resonance: A neuroscience perspective. In Decety and Ickes (2009). [39]

H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. [33]

L. Wispé. History of the concept of empathy. In Eisenberg and Strayer (1990). [36]

Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1972. Translated by G.E.M. Anscombe. First published in 1953. [5, 66]