# Chapter 14

# Psychoanalysis, representation, and neuroscience

The Freudian unconscious and the Bayesian brain

Jim Hopkins

## Abstract

This chapter presents a philosophical synthesis of a range of work relating psychoanalysis and neuroscience. The overall argument is (1) material in these and related fields can usefully be integrated via the notion of representation; (2) the appropriate notion of representation is a biological one common both to psychoanalysis and the Holmholtz/Bayes tradition in neuroscience; and consequently (3) the Freudian unconscious may be understood as realized in what is now described as the Bayesian brain.

Psychoanalysis began as an extension of the intuitive commonsense psychology in whose terms we make sense of one another in everyday life. In this we understand one another as persons who have subjective experience of an objective world, and whose actions are informed by mental states and processes such as those of emotion, desire, belief, intention, and will. Such states and processes have intentionality: that is, they are about things and situations in the world, and they are capable of both truth (or accuracy) or falsehood (or inaccuracy).

This 'aboutness' has historically been regarded as the defining feature of mental phenomena. Recent work in philosophy, however, has enabled us to see that the intentionality of the mental can be understood in terms of the notion of representation; and Ruth Garrett Millikan has argued that representation itself can be seen as a biological phenomenon which has evolved for the selective advantages it confers. So on this account the intentionality of mental phenomena can be understood as realized by the biological functioning of the brain.

Such representational functioning is here explicated via the current Helmholtz/Bayes approach to neuroscience developed from work by Geoffrey Hinton and his colleagues. This approach has recently been advanced by Karl Friston in such a way as to unify a number of basic theories in neuroscience and to integrate them with evolution. This framework is particularly relevant for psychoanalysis, since Freud was trained in the school of Helmholtz and framed his early theories in terms of free energy; and this has enabled Friston and Richard Carhart-Harris to relate Freud's claims to data from neuropsychology, neuroimaging, and psychopharmacology.

The same framework can also be used to link Freud's theories with the findings of affective neuroscience, including the homeostatic regulatory emotional and motivational systems described by Jaak Panksepp and Antonio Damasio, and the interoceptive system delineated by A. D. (Bud) Craig, and in conjunction with research in attachment and development. In this way we can start to describe the wishfulfilling Freudian unconscious as part of the predictive and error-minimizing working of the computational Bayesian brain.

*Keywords*: conflict, aggression, infancy, attachment, repression, interoception, wishfulfilment, superego

## 1. Psychoanalysis, commonsense psychology, and the intentionality of the mental[1]

Psychoanalysis began as an extension of the natural and intuitive commonsense psychology, sometimes now described as *theory of mind* (Wellman, 1990), in whose terms we articulate thought and feeling in daily life. This way of thinking is interwoven with language, and encompasses the motives most basic to our understanding of ourselves. In using it we tacitly but systematically think of one another as *persons* who are *aware* of acting on *mental states or processes* such as seeing, believing, hoping, fearing, and desiring.. This allows us to understand one another with remarkable assurance and accuracy, and so to co-operate and co-ordinate our actions in the far-reaching way characteristic of our species.

To take a simple example: we might understand a person's moving her arm as if to reach a glass in front of her, as an *intentional action* of *trying to get a drink*. In this we would regard her movement as directed by a *desire* that she get a drink and a *belief* that if she reaches for that glass she will do so. Such mental states and processes are said to have *intentionality*, or *intentional content*. (Siewert, 2008). That is, they are *about* things: they *present things to us*, and in a way many have thought that nothing else (including the material brain) could possibly do. And in this they can *accord* or *fail to accord* with the things they are about, in the sense that they can be true or false, accurate or inaccurate, satisfied or unsatisfied, etc.

Accordingly if we say that someone sees, and therefore believes, that there is a glass of water in front of her, we describe both her seeing and believing as having *the same intentional content*, namely *that there is a glass of water in front of her*. This description entails (i) that the seeing and believing are both *about* that glass of water and its location in relation to her, and (ii) that these states would *accord with* the things they are about, in the sense that the perception would be *veridical*, and the belief would be *true*, in the same worldly circumstances. (That is, just if the glass of water they are both about really is in front of the person they both are about, as she sees and believes.)

As this illustrates, we implement such intentional understanding by using language in a particular way. Roughly, we can say that a person perceives, desires, believes, thinks, hopes, fears, remembers, etc., *that P*, where we replace 'P' by a sentence—such as 'there is a glass of water in front of her'—which describes the worldly things or situations which the mental state or process is *about*, and the worldly circumstances in which mental state or process would *accord with them* in the sense of being true, satisfied, fulfilled, gratified, etc. So in this employment of language we shift effortlessly from describing things in the world (e.g. the glass above) to describing states and processes in persons' minds (the agent's experience of seeing the glass and forming a belief about it). We simply *re-use* the words and sentences by which we describe worldly things such as glasses of water and their location in space, to describe persons' minds as directed on, or engaged with, these same worldly things.

These forms of engagement, in turn, are shown as forms of accord or lack of it (truth or falsity, satisfaction or non-satisfaction, fulfilment or non-fulfilment) between persons' mental states and world. So for example when we describe someone's desire as satisfied or her intention as fulfilled, we describe the world as *in accord with* how she wanted or intended it to be. Ultimately we can see these forms of accord as variants of the idea of truth or accuracy; for they obtain when the 'P' sentence in terms of which we describe a mental state is or becomes true. (When someone *satisfies* her desire that she get a drink, the sentence 'she gets a drink' is thereby *made true*. The satisfying act brings the desire into a form of accord—satisfaction—with the world.)

Human beings in all cultures use this mode of description. Like the capacity for language via which it is articulated, it seems, in Chomsky's phrase, 'to grow in the mind'. From shortly after their first uses of words and sentences, children start to embed these in phrases governed by terms such as 'want'. This enables them to make more explicit the way their states of mind mirror the sentences of their language, and so to think and interact with others more efficiently. And in this way they

continue their transition from describing the world to describing themselves as persons with minds who relate to this world in their own individual ways.

Together with this, and perhaps surprisingly, they also begin tacitly to track the most significant *causal relations* which hold between their minds and the world (Hopkins, 1982, 1996, 1999a). Thus we intuitively know, for example, that motives like desire prompt us to act. But we also tacitly specify *how* our desires will cause us to act when we describe them in our commonsense way in terms of their intentional content.

Thus when we say that someone wants *[that she gets] a drink*, we tacitly *describe a cause* [her desire] *in terms of its predicted effects*. For by saying this we indicate the effects that her desire *will have* if she acts on it, both in relation to her *future bodily movements* [drinking movements are predicted] and also *in relation to worldly objects in the immediate environment, with which such movements might engage* [those like nearby glasses, whose manipulation might seem a means of getting a drink]. And we constantly improve such predictions as we consider how agents amplify their desires to incorporate their beliefs. Thus suppose our agent naturally amplifies her desire under the impact of her present visual experience so as to report that she *wants [that she get] a drink from that glass.* Then she and we automatically sharpen our prediction of her forthcoming movements accordingly.

In this way, and without explicit use of the notion of cause, our *that P* way of describing mental states naturally and systematically encodes predictive information about their causal role. And in addition the relations to the environment which we tacitly track in this way reach into our very selves. For when our agent *satisfies* her desire—as marked by her making the sentence 'I get a drink' *true*—she thereby engages the deepest and most encompassing of physiological processes: those of *homeostasis*, by which the brain maintains the physical equilibrium between its body and the rest of the environment which facilitates the basic chemical processes of life.

## 2. First-person authority and emotional conflict

In using language this way, moreover, we find a striking asymmetry, as between our own case and that of others. We ordinarily arrive at a description of the mental states of others by perceiving their bodily activities, as when we see another start to reach out, and understand her as intending to get a drink. In our own case, by contrast, there is no such reliance on external perception. We simply know the relevant descriptions that apply to ourselves, or can bring them to mind, should we be concerned to articulate what we think, want, or feel. If I want a drink, or believe I can get one by reaching for a glass, then I can also think that this is so, and in the same words as I would use to express it to others. This form of self-awareness—the ability spontaneously to understand one's own mental states and put them into words—is often referred to as *first-person authority* (Gertler, 2011). It is basic to our role as agents who can think about alternative courses of action in light of our desires, and so choose how to act; and also to the coordination of each person's actions with those of others, as effected by language, which is characteristic of our species.

Insofar as people can express their mental states in words we can gather the full details of their thoughts and feelings from what they say. If, wanting a drink, I can utter 'I want a drink', then others

can learn what these words mean on my lips, as well as that I am accurate and sincere in my use of them, by matching them with my actions in getting a drink, as well as with my other utterances and actions in other contexts. In this, moreover, we proceed by intuitive interpretive observation, taking nothing just on trust. We tacitly test our tentative understanding of what people say by comparing their utterances with their non-verbal actions (those that 'speak louder than words'), so as to determine the operational significance of words by the way those who utter them act in other ways (Hopkins, 1999a,b).

Again, insofar as we can establish such an ongoing understanding of the language and actions of others, then our knowledge of their motives is limited mainly by their ability and willingness to express these motives to us—that is, by the scope of their own first-person authority. But we constantly and tacitly test this as well, by seeing how far what others say about their motives matches the way we would independently explain their actions. In this way, and in a benign circle, we test others' self-expressive first-person authority, even as we use it as a basis for understanding them.

In order to understand his patients as fully as possible Freud maximized such evidentially valuable first-person expression on their part. He did this by having them engage in *free association*. In this they sought to relax and describe the ongoing, rapidly changing, series of their thoughts and feelings, as fully as they could manage, and without omission or censorship. Such full, far-ranging, and informative disclosure was without precedent in previous psychological investigations, and remains without parallel in other forms even now. This enabled Freud to learn as much about the motives of those he was analysing, as they were able to put into words, and to learn yet more from the patterns that emerged in their spontaneous free associations and other expressive behaviour. Among the things he learned was that there were ranges of motive and circumstance for which first-person authority systematically failed—for which there was good and often-repeated reason to ascribe particular motives to explain what people said and did, and also to explain the forms of distress that had brought them to psychoanalytic therapy; but in which their motives remained *unconscious*, in the sense that they were unable to avow them or think or be aware of them as their own.

Briefly, Freud found that first-person authority was liable to fail when persons were in *deep emotional conflict*, apparently feeling both affection and love, but also hatred and fear, towards one and the same person, characteristically a parent. These conflicts, in turn, were rooted in disparate images of their parents as in relation to them, which Freud called 'the earliest parental imagos' (1933, p. 54); and these went with disparate representations of themselves in relation to their parents as well. In this, as we shall see, Freud anticipated recent work on attachment, which suggests that such representations of the parents have begun to assume a potentially life-influencing form by four months of age (Beebe et al., 2010).

In one set of representations, according to Freud, parental figures appeared as good (comforting, nurturing, helpful, etc.) and as engaging affection, devotion, and cooperation—indeed as 'the prototype' of later relations of love. In another, and by contrast, they appeared very bad (punitive, malicious, cruel, and frightening), and as evoking rage and fear. Since these representations and the emotions to which they gave rise were contradictory, both psychological coherence and family

cooperation required that in general the former set become dominant, and the latter recessive, in the governance of behaviour. This was effected by excluding one set from conscious awareness and first-person avowal, and so from a full role in thought and choice.

Despite this segregation, the early imagos and motives remained active, and expressed themselves in formations that were unchosen and apparently senseless or irrational. These included dreams, bungled actions, and symptoms of mental disorder; and also patterns of feeling and action which were unwanted and self-destructive, as when aggression rooted in the split-off imagos was felt towards others, sabotaging projects and relationships, or again was internalized and turned against the self, as in the ferocious self-criticism which can prompt depression or suicide. Also, and despite their exclusion from reflective awareness, these images were systematically linked with the contents of free association, and liable to be aroused and to become directed towards Freud himself, in what he called *transference* from their original objects.

This enabled Freud to assist his patients in putting the split-off images and motives into words, both as regards the past as they remembered it and also as they felt in relation to his as therapist in the present. As a result his patients were able to *extend* their first-person authority, so as to acknowledge these images and motives as their own. They could thus seek to integrate the conflict-engendering representations by working them through in feeling and in thought, and thereby modifying and ameliorating them. Such use of insight and feeling to mitigate conflict via the extension of first-person authority remains the hallmark of psychoanalytic psychotherapy.

## 3. Intentionality, mental representation, and the brain

Freud's basic explanatory concepts thus included the intentional notions of commonsense psychology, as well as others, such as *wishfulfilment*,[2] discussed below. How do such interpretive mental and intentional notions relate to cognitive science, neuroscience, and other recent approaches to the mind? We can see this more clearly, and move towards integrating these fields, by focusing on the notion of *representation*—as used, for example, in connection with the Freudian term 'imago' in the paragraphs above.

We have seen that mental states are related to the world in a way comparable with the sentences we use to describe them. Thus both the belief that there is a glass of water in front of me and the sentence 'There is a glass of water in front of me' are about that glass and its relation to me, and both are true if that same glass of is water in front of me (and false otherwise). The words and sentences, however, are paradigmatic instances of *linguistic representations*. We can, for example, classify them together with maps, blueprints, photographs, drawings, etc. Just as a blueprint, drawing, photograph, or map may represent the spatial locations of objects, so does the sentence 'There is a glass of water in front of me'. When we say that a picture is worth a thousand words, we are comparing the efficacy of different kinds of representation. And all these kinds of representation are said to be about the things they depict, and to be capable of accord in the sense of truth or accuracy.

But then if representations are marked by the same features as the states to which we ascribe intentionality—if it is in the nature of representations to be about the things they represent, and to

be capable of truth or accuracy— it appears that our mental states may have intentionality simply because they are, or embody, forms of representation. This is the *representational theory of mind* (Pitt, 2008), advocated by many philosophers, psychologists, and cognitive scientists (see, e.g. Horgan and Teinson, 1999; Thagard, 2010). On this account in saying that a person believes that there is a glass of water in front of her, we describe her as *mentally representing* how things are; as in saying that she desires that she drink from that glass, we describe her as mentally representing how she wants things to be. The ability mentally to represent things, in turn, seems to imply the use of some kind of internal system of representation, as realized in the brain.[3]

Ruth Garett Millikan (1982, 2005) has recently explicated the notion of representation in biological terms, as precipitated during the course of evolution for the selective advantages it confers. This provides an account of the general notion of biological information (Godfrey-Smith and Sterelny, 2008) to which many accounts of biological regulation refer, and can be applied even to the working of DNA (Shea, 2007). In the light of such accounts we can proceed here on the assumption that all the forms of representation with which we are concerned—including those of commonsense psychology, cognitive science, psychoanalysis, and neuroscience—devolve from those that evolution has conferred upon the brain.[4]

## 4. Representation in commonsense psychology and computational neuroscience (the hierarchical Bayesian brain)

We can see more about these powers by briefly considering a simplified account of the Helmholtz/Bayes neurocomputational framework now referred to as the 'Bayesian brain' (Doya et al., 2007). This began to emerge in its contemporary form in 1995, in work in which Geoffrey Hinton and a number of his colleagues sought to use an information-theoretic account of free energy to embody basic insights of the account of perception put forward by von Helmholtz in the nineteenth century. (Dayan et al., 1995; Hinton et al., 1995).

In a series of recent publications Karl Friston (2003, 2007, 2010a,b) has sought to use this Helmholtz/Bayes framework in a number of closely related ways. First, to unify basic theories in neuroscience (Friston, 2010a); secondly, to integrate neuroscience with the theory of evolution via the mathematical relation of free energy to thermodynamics (e.g. Friston and Stephan, 2007); and third, with Richard Carhart-Harris, to present Freud's theories, which Freud himself had framed in terms of free energy, as consilient with a large range of data drawn from neuropsychology, neuroimaging, and psychopharmacology.

In this framework the brain is taken as a predictive 'inference or Helmholtz machine', which uses hierarchical Bayesian inference so as to extract statistical explanatory patterns from data of all kinds, and in a maximally efficient (optimal) way. The Bayesian inferences are realized in hierarchies of representation-producing neural networks which cooperate to 'optimize their representation of the sensorium' by constructing 'top-down prior [Bayesian] expectations about sensory samples from the world' (Carhart-Harris and Friston, **2010).**

## 5. Philosophical antecedents of the Helmholtz/Bayes approach

Helmholtz wrote in a tradition founded by Immanuel Kant (1781). His neuroscientific work partly embodies Kant's idea that we can see our basic concepts (Laurence and Margolis, 2011)—that is, our basic but everyday ways of thinking of space, time, substance, objects, events, and the relation of cause and effect—as performing an *unconscious synthesis* of the 'manifold' of sensory intuition. According to Kant this synthesis transforms the sensory manifold into *our manifest conscious image of ourselves as self-aware subjects of experience which is internal to our minds, but which we understand as caused by objects and events in the world external to our minds, which objects and events are in causal relations with one another, and so with our bodies and sensory organs when we perceive them*. This philosophical perspective, at once straightforward and profound, has been carried forward via Helmholtz, Hinton, Friston, and others, into the conception of the Bayesian brain. For according to this account the brain uses our concepts in working top-down from what Fuster (2009, p. 2047) calls 'the highest [conceptual] levels of association cortex' so as to represent sensory neural input in terms of '*the causes of exteroceptive and interoceptive sensations*' (Carhart-Harris and Friston (2010).

This latter terminology was introduced by Sherrington (1906) to describe input to the nervous system by the sources of its sensory origin. So (roughly) *exteroceptive* input concerns what Kant called *outer sense*: the detection of light and sound, as by sight and hearing (but also detection of heat via the skin, or non-auditory detection of vibration): of chemicals by odour or their capacity to irritate the skin; of shape and texture by response to pressure or touch, and so forth. Turning to what Kant called *inner sense*, *proprioceptive* input comes from the networks which register and control the position and motion of the limbs. And *interoceptive* input, as discussed in recent work by Craig (2009, 2010), apparently comes from a range of sources inside the body, including C-fibres and networks relating to subcortical mechanisms of homeostasis and emotion. These seem to map via the hypothalamus and thalamus to internal topographic body maps in insular cortex, so as to yield a range of sensations and experiences about how we feel inside. Although this pioneering work will be modified by further observation and theory, it provides a model which we can use in what follows.

## 6. The central role of sensory input

Now it may seem remarkable that a comprehensive account of the working of the brain should (like Kant's account of unconscious conceptual processing) be focused on sensory input as opposed to motor output. But the central role of input seems established by a simple contrast put forward by Damasio (1999) and elaborated in more detail elsewhere (Parvizi and Damasio, 2001, 2003). The main lines of sensory input (exteroceptive, proprioceptive, and interoceptive) are adjacent in the brainstem to those which carry motor output. Damage to the output lines produces the pervasive paralysis of 'locked in' syndrome, in which thought and conscious experience remain intact. Damage to the input lines, by contrast, causes consciousness- and mind-obliterating coma. So it is (broadly) sensory input which drives the basic sense-making and experience- and consciousness-producing operations with which we are concerned. (And in the Helmholtz/Bayes conception motor output is accordingly taken as governed via the sensory input to which it gives rise.)

## 7. Fitting neuroscience with common sense in representing sensory input as experience of its causes

We can see these idea at work if we return the thirsty agent we have been imagining from the start, who has seen a glass of water and is intending to drink (from) it. We can now observe two things:

(i) We can fit our commonsense forms of description together with the descriptive terms from Sherrington's neuroscience by noting that this agent is using her system of concepts (including those of the self, space, objects such as glasses, etc.,) to understand her current *exteroceptive* (visual) and proprioceptive neural input (from the fibres innervating her muscles, tendons, joints, inner ear, etc.) as *awareness of seeing a glass of water located in front of her body and within reach of her hand*.

(ii) Likewise (and we will discuss this further below) we can also say that she is understanding her current *interoceptive* neural input (from inside her own body, as above) as *awareness of her own thirst (an internal and homeostasis-related cause itself caused and predicted by hypothalamic activity) and awareness of her own desire to drink (a potential and predictive cause of future action as caused by thirst)*.

That is: insofar as we regard someone as an agent who is conscious of herself in the world—and hence aware of her own motives and environment—we can also regard her brain as *representing its own neural input* as *caused by the external objects and internal states (motives such as thirst, belief, and desire) of which she is aware*, and *thereby constituting her as a person in the sense of a Kantian subject of internal experience who is an agent acting in the external world*.[5]

We said at the outset that we naturally think of the mind as an inner realm of experiences and other mental states and processes which have the remarkable feature of intentionality. We can now see (at least on the Helmoltz/Bayes conception) that we do this precisely because this is the image of ourselves that our brains create for us. The intentionality of the mind, and the mind's (or brain's) image of itself as a self-aware subject and agent in the world, are realized via the representational powers of the brain. And since this physically produced image includes the inner causes that we regard as reasons, emotions, or motives of other kinds, the working of such motives, and their compatibility with physical causality, should be regarded as integral to our image of ourselves.

## 8. Hierarchical Bayesian representation

We can get an idea as to how the brain might represent input as caused in this way by regarding the representation as generated by hierarchies of neural networks working in the ways described by Hinton and Friston. These would project forward from those which first receive sensory input, through a series of intermediate levels which are progressively more encompassing and integrative, up to those which finally realize person-level concepts, beliefs, and desires. (Thus networks at the sensory peripheries would project forward to intermediary levels, which also project backwards and sideways; networks in the thalamus, perhaps, to others in sensory cortex, which also project backwards and sideways; those in primary sensory to others at a higher level; and so

on.) The forward projections are often topographic and drive processing forward and up the levels; the backward are modulating or inhibitory in relation to those below.

In this conception, assemblies or networks at each level work simultaneously, both from the top down in relation to those below them, and from the bottom up in relation to those above. That is, each works top-down:

(i) To produce accurate predictive representations of activity at the level driving forward from below, so as

(ii) To send these representations back as predictions which *inhibit and modulate* what is happening below. The effect of this is partly to suppress predicted activity, and also, by leaving an *uninhibited* remainder, to specify and magnify the effect of *the unpredicted part* of the activity driving forward from below.

And also bottom-up:

(iii) To send forward enhanced and specific representations of *errors in prediction* in the suppressive representations sent back by the producers above, so as

(iv) To cause those producers to send *updated and better* suppressing and specifying predictions back again.[6]

In all this the accuracy of the predictive representations at each level consists simply in their unfolding over time so as to match those unfolding at the level below. This is continuously tested, by all higher-level producers sending current representations back along inhibitory connections to suppress what they match below. The remaining forward-driving representational activity perforce further specifies the errors in the original prediction; and this goes forward to the level above to cause the production of a better match. In this way all representations improve their predictive scope and accuracy at every turn.

This architecture is consistent with the 'massive scaffolding of hierarchically organized memory networks in a continuum of increasing network size from the primary cortex to the highest levels of association cortex' described independently in Fuster (2009, p.2047), and seems to fit with other data so far accumulated about the brain. As Hinton and Friston have stressed, such a computational structure can extract all the probabilities it requires for building its constantly self-improving hierarchies of representation empirically from its initial sensory inputs. And conceptually speaking, it seems that such a process could enable our agent's brain to provide her with the subjective experiences of *thirst* and *seeing a glass in front of her* that we take her to have.

For as we have said, the most encompassing and integrating representations in her brain would be those realizing the personal level representations of conceptually informed consciousness and belief. In this each of her concepts (*thirst*, *desire*, *drink*, *sight*, *belief*, *glass*, *water*, etc.) contributes to making her experience intelligible as it occurs, and also in yielding further predictions about how it will unfold. For example her regarding the glass she sees as transparent but solid, and the water within as also transparent but liquid, makes sense of how they look together, move in relation to one another, and so on. And this appearance also predicts to her that she will be able to grasp the

glass and pour the liquid into her mouth if she moves *like so*, as her thirst and unfolding desire and intention to drink will shortly prompt her to do.

So as far as we can see by considering the working of our concepts, inhibitory predictions returning from conceptual high level might indeed be reaching an equilibrium with input coming forward from the external senses and inside the body in the way described. The meeting place, in Kantian terms, would be that at which my (downgoing and side-by-side connected) *concepts* meet my (upcoming) *intuitions* so that the latter can be understood as *my own (inner, subjective, and private) experiences of the (external, public, and objective) world I share with others*. By applying concepts from above at the same time as carrying upwards sensory input from below, such a hierarchy of self-correcting representations might indeed impose on neural input a continuous relaxing top-down person-level predictive and error-minimized conceptual representation of experience of its external and internal causes, conceiving these as internal states and external objects of which we are aware.

In this what Freud called 'psychic acquisition'—the whole generative predictive model of the world we have attained in experience—would, as he and other early neuroscientists supposed, reside in the (backward, modulating, inhibitory) connections between neurons over the interlinked representation-producing hierarchies of the brain.[7] And because the initial parameters of such producers are set by evolution, the model of the world they compose on the basis of sensory input will likely depict it as emotionally significant, motivationally engaging, and presenting opportunities for life-sustaining activity (Friston, 2003, 2010a,b).[8] Also, by providing an account of the brain as producing *subjective experience of the self as experiencing an objective world*, the account serves to fill out previous discussions of consciousness,[9] and materials to resolve long-standing philosophical problems about consciousness as well.[10]

## 9. An example of conflict in visual input

We can plausibly see both the top-down and bottom-up working of these hierarchies in experiments with artificially induced binocular rivalry. These arrange for the right and left eyes to be given visual input depicting different objects, for example a face on the one hand and a house on the other. In such a situation the experiencing subject oscillates between these alternatives, seeing a face, and then a house, and then a face again, and so on, with elements of one alternative sometimes 'breaking through' before it dominates and the cycle goes on. As Hohwy et al. (2008) have argued, this is how we should expect visual experience in these circumstances to be, if the brain was working at higher levels to represent the input to the eyes as visual experience caused by the objects, that on a Bayesian account would be most likely to have done so.

A brain representing input as caused by a house would not also represent that input as caused by a face. Since we never experience faces and houses (or other distinct types of material objects) in the same place at the same time, high-level representation-producers implementing the belief-and experience- informing concepts *face* and *house* would antecedently set the probability of seeing a thing which gave sensory input as simultaneously from both as nil, as is reflected in our inability to vizualize such a thing. These concepts, therefore, will have a strong inbuilt relation of mutual

disconfirmation or inhibition, so that initial top-down representation will be of the input as caused by one or the other, but nothing like both. Suppose the concept *face* initially dominates. Then the overall input will activate the concept *face*; the visual input will be represented as caused by a face; the experience will be one as of seeing a face; and modulatory and/or inhibitory predictions about probable input caused by a face will be sent back along the hierarchies for matching with the actual input coming forward from the eyes.

On Bayesian calculations success in matching will raise the posterior probability that the input was caused by a face, and failure (= prediction error, or again free-energy) will lower it. In this case matching must fail, because no representation of the input as caused by a face will match that coming forward from the house-stimulated eye. Since this failure is uncorrectable—the input from the house-stimulated eye is veridical, and so cannot be explained as coming from perception of a face, and so cannot be 'explained away' by top-down use of the concept *face*—this failure will lower the posterior probability that the input is caused by a face so as to effect inhibition of the concept *face* and prompt activation of another.

As before, the only concept that would be capable of matching the input—that of something which produces sensations as of seeing a face and house in the same place at the same time—will be ruled out as antecedently improbable and visually unrepresentable (even if it breaks through for a moment as tentative best hypothetical explanation). So now the concept *house* will dominate; the experience will be that of seeing a house; and the cycle will go on, as it is observed to do.[11]

## 10. Conflict more generally

This illustrates the capacity of the Bayesian approach to provide a compelling account of the subjective content of experience; and at the same time it shows *the remarkable ease with which the brain can alter consciousness so as to remove the effects of ongoing and veridical sensory input*, in a process that could be taken as akin to Freudian repression. The experiment suggest that what is required for this is simply that *the input to be suppressed/repressed is inconsistent with the dominant model that the brain is currently using to make sense of experience.*

In such Bayesian processes, as we recall, the brain is

(i) 'optimiz[ing]. . . representation of the sensorium', by

(ii) 'constructing 'top-down prior expectations about sensory samples from the world", by which the brain

(iii) represents 'the causes of exteroceptive and interoceptive sensations'.[12]

With this in mind let us see how the example above can be understood in terms of *managing conflict*, and in a way which links it with Freud. We can do this in a series of stages, by describing

(i) Conflict in current perception, but introducing Freudian terms for Bayesian functions as suggested by Berlin and Koch (2009); and then discussing

(ii) Conflict among motives and emotions; and then

(iii) Conflict of the kind described as Freudian in 'First-person authority and emotional conflict' above; and so finally

(iv) The two basic kinds of Freudian conflict (as between motives directed at a single individual, internalized as conflict between parts or aspects of the self) illustrated in an example from Freud.

(In the space available these sketches must perforce be brief and incomplete.)

## 11. Binocular rivalry as concept-driven conflict on two levels

The first step is to observe that the example of binocular rivalry we have just considered can be taken as representing how the Bayesian brain works top-down as well as bottom up in seeking to manage conflict. From the top, the brain is seeking to impose *incompatible (conflicting) concepts* (*face*, *house*) on visual input, where this incompatibility is a product of prior assumptions in the underlying generative model. The use of these incompatible concepts, moreover, represents the *real* and *veridical* sensory inputs involved in the experiment as *incompatible with one another*, in the pragmatic sense that (despite their veridicality) each is bound to be treated as error while the other is represented in consciousness as experience of its cause. In consequence, the input kept split off from consciousness is *thereby also kept insistent and active while it remains unconscious,* in virtue of its role as error signal that cannot be eradicated.

Input not subject to such conflict, is, as Friston says, 'explained away',[13] by its representation as conscious experience of its cause. This contrast is important, for it indicates that

(i) Such Bayesian conflict-suppressed unconscious sensory input as we find in the face/house case has a role closely analogous to that of material subject to Freudian repression, in the sense that it is perpetuated (as error signal) by being rendered unconscious. This entails that (so long as it is active) the input presses upward for conscious representation, and hence remains in causal and representational conflict with the dominant conceptual model. So also

(ii) In contrast when non-conflictual input is represented in consciousness its 'energy' can be said to be fully 'bound' (put to use in psychic work) as opposed to 'free' (or at least as opposed to *wasting in representational conflict*). And in general, as we shall see, emotional conflict will appear as a source of free energy in the Bayesian as well as the Freudian uses of this notion.

In addition, we can see that the processes which yield our conscious conceptually informed image of the world are themselves unconscious, and their activity is reflected in the manifest image only *after* events they represent have already occurred.[14]

## 12. Bayesian repression; the Bayesian conscious ego; sensory systems preconscious and unconscious, and a Bayesian superego

With this in mind let us follow the lead of Berlin and Koch's (2009) 'Neuroscience meets Psychoanalysis' and substitute 'repression' for 'suppression' where this is appropriate in the Bayesian account. To do this we can:

(i) Introduce the term 'dominant (top-down) conceptual model', as used above, for the (evidently vast and interconnected) set of conceptual representations the underlying generative model is currently employing to explain (away) input by representing it as conscious experience of its cause; and also

(ii) Describe the *veridical* input currently successfully explained by the dominant conceptual model as *accurately conceived (represented, etc.) in conscious experience* (or in consciousness, etc.) *as experience of its cause*. (In the experiment this would apply first to input from the face-stimulated eye being successfully represented by the dominant conceptual model as visual experience of seeing a face; then to input from the house-stimulated eye being successfully represented by the dominant conceptual model as visual experience of seeing a house, and so on). This allows us to specify a kind of

### 12(a). General Bayesian repression

For now we can describe as *repressed* and *rendered unconscious* all *veridical* sensory input which is accommodated by some conceptual (or proto-conceptual) model but kept from consciousness via conflict with the conceptual model which is dominant overall. This input will be *repressed together with its accommodating models*, as in the face/house case. (So while the concept *face* was dominant this would apply to the input to the house-stimulated eye, as accommodated by the concept *house* now in this unconscious role; and vice-versa as the concept *house* became dominant so that the conscious experience became that of seeing a house.)

In this we do justice to the fact that material which is repressed (in this Bayesian sense) because of conflict between partly veridical models is not just veridical for the models that accommodate it. Rather, as the experiment illustrates, it may also be *potentially veridical current conscious experience for the subject concerned*. For if input is veridical at least as accommodated and repressed, there may also be a concept on which it would be a veridical part of the dominant model—if only the brain could frame or use this concept, as in the face/house example it cannot do. So overall this gives

### 12(b). A coherent Bayesian conscious ego, inhibiting and modulating downwards and also interacting with subpersonal systems which deliver information

This in turn means that we can roughly but reasonably regard the dominant conceptual model as constituting a *conscious ego*. For:

(i) The dominant conceptual model, like the Freudian *Ich*, continuously determines the conscious experience of the subject, both as regards awareness of objects in her external environment and also as regards awareness of her own internal states of mind. (That is: according to the Helmoltz/Bayes account, this set of currently cohering conceptual models really is now producing in each of us our overall conscious image of ourselves in the world.[15]) In addition:

(ii) This ego, as we will suggest in more detail later, is continually *repressing and keeping unconscious* both veridical neural input and veridical models of ourselves which accommodate this input, but which are in conflict with the dominant model. (For again in the face/house case the conscious representation of a house [face] is repressed and rendered unconscious, not because

the veridical house-input [face-input] from the sensorium has ceased pressing upward, seeking expression in consciousness; but rather because use of the concept house [face] which accommodates this input has itself been repressed or inhibited.) Again, as we have seen, repression of this (Bayesian) kind must serve to keep the repressed sensory input and accommodating models alive and seeking expression, as this is underwritten by its role as error-signal that cannot be explained away. (And by now such repression of what does not fit the dominant conceptual model may start seriously to remind us of Freud's claim discussed above in 'First-person authority and emotional conflict', that the 'earliest parental imagos' become recessive but also remain active, owing to their banishment from consciousness.) Also

(iii) This ego should interact in a holistic way with *subpersonal* representation-processing mechanisms—those whose workings are *not* such as to enter consciousness, but nonetheless inform it. Thus we have subpersonal neural mechanisms which enable us to make and hear *sequences of sounds* as *utterances* of *sentences expressing thoughts, hopes, desires, threats, etc.* Such mechanisms are studied in various ways throughout the mental sciences, and their operations encompass other executive functions that Freud assigned to the ego. Finally

(iv) As well as interacting with subpersonal mechanisms this ego will interact with *other currently repressed person-level conceptual representations*, such as those primed by current experience and ready for interpreting what is in the offing, as well as others relatively remote from predicted experience but ready to enter if required (e.g. to represent a barely noticed movement in the shadows, or an approaching figure, as the dangerous predator it may turn out to be.)

Overall these observations suggest that we can regard such an ego's effectiveness and cohesion as depending on the adequacy and coherence of the agent's (or the underlying generative model's) *system of concepts*. The elements of this will be able to activate, deactivate, and interanimate one another to produce experiences, beliefs, etc. in a side-by-side, cooperating, and holistic way, as the same time as each does its top-down work, again in cooperation with others, in using such beliefs to explain (away), and in this way to bind, input pressing up from the sensorium. Also we should expect the work of producing and using belief (and desire) in this way to be integrated with the use of memory (working and long-term, perceptual and executive) as described in Fuster (2009).[16]

**12(c) Preconscious and unconscious systems, and a Bayesian superego**

We have so far considered two forms of person-level but repressed/suppressed unconscious functioning:

(i) That of conceptual models ready to enter consciousness as required by perceptual input; and

(ii) That of conceptual models accommodating one or another kind of veridical sensory input, but which are repressed because they conflict with the dominant conceptual model (as in the face/house case, or again those discussed by Freud).

Of these (i) gives us a Bayesian preconscious and (ii) a Bayesian repressed unconscious, which has an overall causal structure strikingly similar to that delineated by Freud. And we can take a further step in Freudian rephrasing by considering another feature of the face/house example.

What keeps the Bayesian ego in the perpetual conflict-driven oscillation we observe in this experiment? After all, the ego (or the underlying conceptual system or generative model) seems capable of *creating* a concept which would explain the novel input and resolve the conflict—that is, a concept of some sort of thing which produces sensory input as of seeing a face and house in the same place at the same time. For this seems to have been the concept the brain was trying to use—in the 'breakthrough' experiences described by participants—to explain the novel input it was confronted with; and (rescinding from impossibilities in visualization) this might have provided an accurate account of the cause of visual experience in the strange but real experimental set-up in which it actually found itself. We have implicitly been celebrating the Bayesian hierarchies as paradigms of powerful learning from experience: but in this case the system—in receipt of input that was both constant and veridical—remained a resolute non-learner. Rather it seems repeatedly to have strangled the new more predictively adequate idea each time it emerged, so as to go on with its now thoroughly discredited cycle of conflict-driven repression, using the now provably inadequate concepts *face* and *house* ...

Here we can say that the underlying generative model, at the same time as acting as ego, was also acting as a kind of

**12(d) conservative Bayesian conceptual superego**,[17] whose insistence on adherence to prior modes of thought prevented (it in its role as) the ego from employing a concept framed for this new case. (And although on the Helmoltz/Bayes account the brain works empirically with input from the beginning, such conservative conceptual favourites may enjoy legacy admission to consciousness on the basis of ancestral inheritance.)

So we can say further that in this case the ego is caught between two masters, the conceptually conservative superego and the sensory id. Unprecedented but veridical face- and house-stimulated input is pressing up from the id, seeking expression in consciousness. The concepts the ego naturally first employs for this are inadequate and incompatible, so that the ego perforce continually finds itself repressing veridical input, which therefore continues pressing upward, in a way that might be contained but must remain dynamically active. The ego itself might respond to the input in a way which meets its novelty, that is, by framing a concept of a sensory cause which would enable it to represent the input in consciousness, and so to end the cycle of conflict which the use of prior concepts generates. But the use of such a novel concept is continually aborted by the inbuilt conceptual conservatism of the superego. So the ego, divided against itself, oscillates in producing alternative conflicting states of conscious experience.

This too has a certain fit with Freudian concepts: so perhaps something of this part of the Freudian picture is also Bayesian, even in the example we have been examining.

## 13. Requirements of a Freudian model

Now of course even if such redescriptions can be made to reproduce Freud's own, they fall far short of yielding a genuinely Freudian model. The face/house example may instantiate notions of conflict and repression, but it plainly lacks the core Freudian features of *emotional* conflict and *long-term but repressed and active experiential autobiographical memory*. Still it provides an

account of something which seems akin to hysterical or hypnotically induced sensory illusion or blindness; and Berlin and Koch (2009) suggested the use of 'repression' in binocular rivalry precisely to facilitate comparison with cases of this kind. Thus they cite the patient described in 'Blind and sighted in one person', by Waldvogel et al. (2007). This patient, who suffered from dissociative identity disorder, was originally diagnosed with cortical blindness. She recovered sight after 15 years of psychotherapy. This must have focused on the severe emotional conflicts characteristic of this disorder, which are shown in the dissociations which constitute its symptoms. Her step-by-step recovery, moreover, permitted comparative electroencephalographic (EEG) evaluation of alternating blind and sighted states.

This indicated that while blind the patient maintained greatly reduced activity in her primary visual cortices—even while facing input to her eyes which caused readily detectable cortical activation when she was sighted. As Berlin and Koch report, there is no known mechanism by which such an effect could be consciously produced by a subject with open eyes and capable of sight. So this finding (like many others less clearly documented) seems to imply that the brain can intervene at an early stage to suppress visual input, even before it reaches visual cortex. At present there seems no better explanation for such findings than a Bayesian process of the kind we have been describing, but which turns on unconscious emotional conflict of the kind delineated by Freud.[18] So how should we extend our discussion to take such conflict into account?

## 14. Assigning a fuller role to interoceptive input

As already noted, the sources of sensory input which the brain represents as awareness of motives such as thirst and a desire to drink would appear to be those of the interoceptive system, as recently delineated by Craig (see 2009, 2010). In light of this we can follow Solms and Turnbull (2002) in describing the 'inner world' of the interoceptive sensorium by reference to the empirical tradition of affective neuroscience which includes Panksepp (1998), Damasio (1999), and Damasio et al. (2000). Accordingly the neural inputs we are aware of as various forms of motive or emotion—as thirst, or wanting to explore or play, or as feeling the pain (or panic) of separation, or again as rage or fear or wanting to find something out, would trace back

1. To the hierarchies (which on the present account might also be Bayesian) of the 'multi-tiered and evolutionarily set neural mechanism aimed at maintaining organismic homeostasis' in terms of which Damasio and his colleagues (2000, p1049) conceive both homeostasis and emotion, or again

2. To the 'multiple prototype emotional regulatory systems' which Watt and Panksepp (2009, p. 93) describe as 'sitting over homeostasis proper (hunger, thirst, temperature regulation, pain, etc.)' and 'giving rise to attachment', or again,

3. To the process of attachment itself, which Watt and Panksepp (2009, p. 93) describe as establishing the 'massive regulatory-lynchpin system of the human brain'. This system exercises a 'primary [top-down, which again on the present account might also be Bayesian] influence over the prototype systems below'.

This would allow us to extend the Helmoltz/Bayes approach to the emotional and motivational depths of the limbic and subcortical areas of the brain, and at the same time directly to consider the sources of emotional conflict involving 'the earliest parental imagos' to which Freud assigned a prototypical role. As he stressed, 'the major needs' provide 'endogenous stimuli' which the brain cannot escape (1895, p. 297). Their demands may conflict, in the sense that they cannot be met by the same patterns of activity; and the infant depends entirely upon its carers for their satisfaction. So here we can also bring a long-standing tradition of empirical psychology to our aid. For while the establishment of these early prototypes (or the proto-concepts which embody them) should be regarded as among the first and most basic empirical tasks of the brain, this process has also been studied intensively in the fields of attachment and developmental psychology. These have recorded an important range of experimental and statistical results.

## 15. Interoception, motivation, and free energy

Even in the deep interoceptive cases we are considering, the sensory inputs we experience as motives are characteristically made conscious in terms of feeling and desire. For it is by producing desire that such input in turn produces intentional action aimed at correcting whatever internal disequilibrium —homeostatic or emotional – is producing free energy (= error in prior calculations as to what action would be optimal.)

This is readily illustrated by the thirsty agent we have been considering, who will naturally convert her *depictive* representation of the environment, which shows a glass of water in front of her, into an *action-directing* representation, in the form of a *desire now to reach out to get that glass and drink from it*. Such a representation perforce also *predicts her own forthcoming sensory experience, including the experience of satisfying her own desire* via *the bodily movements leading to and including her drinking water from the glass*. So she then straightway *acts to make these predictions about the course of her own experience come true*, thereby confirming the model of herself in the world on the basis of which these predictions are being formed. (This is the process Friston et al. (2010, p. 6) describe as 'sampl[ing] the world to ensure our predictions become a self-fulfilling prophecy'.[19] But here this appears in the commonsense form of intentional action aimed at the satisfaction of desire.

Such action has two distinct sorts of consequences, which are temporally coordinated:

(i) In the short term, the predicted *experience of satisfaction* (in this case, that of drinking) pacifies the just-generated desire to drink on which the agent has acted, and so suspends the Freudian 'demand for work' embodied in that desire. This allows the agent to turn to other tasks, while the deeper homeostatic adjustment caused by the water she has just taken into her body gets under way.

(ii) In the longer term, the water makes its way into the agent's bloodstream, where it accomplishes the work of restoring the original homeostatic imbalance while the agent's desire relating to this remains pacified.[20]

## 16. Desire and predictive representation

At this stage it may be worth making more explicit how these hypothesized processes fit with the commonsense psychology with which we began. We know that desires are causes of actions which satisfy them, and that they are pacified—caused to cease to operate—by the experience of their satisfaction. So designating our agent by 'A' and abbreviating 'desire' by 'des' and the appropriate causal relations by '→' we can represent the lifecycle of a desire to drink such as we have been discussing as follows.

**17. Phases in the satisfaction and pacification of desire**

A des that A drinks → A drinks → A experiences, believes that A drinks → A's des that A drinks pacified

And since this applies to any desire which prompts satisfying action, we can schematize it in a general way by:

A des P → P → A exps, bels P → A des P pacified

In this artificially simple but schematic representation we find four phases in predictive and causal sequence:

(i) The inception of desire in A des P. This, in the discussion above, reflects the initial working of the brain in representing the internal sensory input caused by a lack of water as an experience of its cause (thirst), and hence as generating a further internal cause, namely a desire for action which will relieve the thirst, and in this way will address the underlying homeostatic imbalance.

(ii) The satisfaction of desire, in A des P → P, in which the agent actually drinks; and the latter as acting

(iii) As in P → A exps, bels P, which represents the *sequence* of believed and veridical experiences (*experiences of satisfaction*) of the agent's satisfaction of her desire. These are the experiences predicted both by the agent's thirst and her desire to drink, which as sensory predictions the agent herself makes come true; and finally,

(iv) A exps, bels P → A des P pacified. This represents the pacification of the agent's desire to drink which follows upon her experiences of quenching her thirst by drinking in (2) and (3) above. According to the exposition here this is the first phase of the Bayesian version of 'explaining away' which applies to an internal cause of experience such as thirst or a desire to drink.[21]

**18. A contrast between external and internal causes of experience**

Thus overall we are placing the perceptual experiences of satisfying desire together with those involved in the formation of beliefs about faces and houses at the highest conceptual levels of the Bayesian hierarchy. As noted, however, there is an important contrast between them. In the formation of beliefs about faces and houses on the basis of sensory input the higher conceptual levels suppress input coming from sensory sources below by predicting their activity as caused by experience of objects *external* to the self. In the formation of beliefs about internal phenomena such as thirst and desire, the conceptual levels likewise represent sensory input as experience of causes, and in this case also endow us with first-person authority about (many of) these causes, as discussed above in 'First-person authority and emotional conflict'. But in the interoceptive case

this is often only the first step in a series which leads through intentional action and the accompanying experiences of satisfaction to corrective alterations in the underlying homeostatic or emotional/motivational processes which are the ultimate sources of the desires with which we are concerned.

Here, on the present account, the wheel comes full circle. In veridical perception of the environment in general the higher personal and conceptual levels suppress (relax) the lower, by successfully predicting their input as caused by, and so as experience of, external objects. In the perception of the self in desire-satisfying action, by contrast, the hitherto lowly inputs from the external senses suppress (relax) the higher levels, by pacifying the person-level desires and intentions which both predict and cause these inputs, while the activities of satisfaction themselves bring deeper homeostatic or motivational changes.

Insofar as this is correct the final units in all hierarchies to be affected in such a cycle of successful action will be those in at the bases of the 'multi-tiered and evolutionarily set' mechanisms for homeostasis and emotion envisaged by Damasio and his colleagues.[22] These subcortical networks—which in this account we can see as psychologically as well as physiologically the most fundamental—are the final targets of the quieting of internal disequilibrium (or error or free-energy) effected by getting a drink that one had previously come to desire. Their silencing marks the recovery of a satisfied mind.

## 19. Attachment and infantile emotion and experience

The central role of attachment—the forming of basic emotional bonds between the infant and its carers, among whom the mother is statistically foremost—is apparent from consideration of the basic (homeostatic, emotional, regulatory, motivational) systems in (1), (2), and (3) above. For these systems enjoy inbuilt relations of excitation and inhibition, and come connected for expression via the newborn's face, voice, and movements. Their early and vigorous activity—for example in a hungry baby's uniquely demanding, distressing, penetrating, and mobilizing cry—is the helpless human infant's main means of directing parental attention to its needs and enforcing investment that will fulfil them and so enable it to thrive.

The ensuing dialectic of demand on the part of the infant and satisfaction (or non-satisfaction) on the part of the mother (or other carers) provides the context of what are arguably the most important experiences of life. These are the early experiences of *the self as relating to others in a context structured by the basic needs and emotions of the self*, which are *as yet unknown by the self*. These experiences shape the infant's cortex (and hence its nascent and growing concepts) as it begins its own process of post-natal development, via critical phases of synaptic growth, myelination, and experience-dependent neural pruning. Such neural development thus coincides with the infant's use of its experience—and particularly experience of its interactions with the investing mother[23]—to start to build representations of its own self and the internal causes of its own behaviour as in relation to the other objects of its experiences and emotions.

In light of this it appears (1) that we should see complex human feelings as rooted in the orchestration of the basic subcortical mechanisms of homeostasis, motivation/emotion, and attachment, as these have become both corticalized and socialized over the evolution of our

familial, articulate, and group-forming species; and (2) that the basic representations fostering this orchestration are achieved via cortical development under the impact of *the infant's early experiences of relationship* and *in contexts first prompted and regulated by these basic subcortical mechanisms*.

For during this early period the mother responds to her baby's expressions as the principal satisfier of homeostatic needs, pacifier of various forms of distress, provider of opportunities to learn, and securer of ease and peace of mind. So she is, for example, the main object of reward-seeking exploration (Panksepp, 1998), and so the main source of the pleasures of liking, the compulsions of wanting, and the experiences of learning (Smith, Berridge and Aldridge, 2011) *as these relate to every source of internal and external sensory input*. Again, she is the first partner in play, proto-conversation, and other pleasurable social interactions, and the first to be missed, yearned for, or grieved.

But then also, in her inevitable shortcomings in such essential respects, this very same mother is *the first easily discernable external candidate for the role of cause of all forms of deprivation and frustration*. Her imperfect timing (or imposition of order or schedule) is the first salient external cause of hunger, or again of the panic of distress at separation, which in early life might well be as felt threatening loss of all resource. So she is also the first object of full-throated rage and deep-seated anxiety and fear, as expressed (perhaps together with distress at separation) in primordial form in a hungry infant's raging cry.

Wittgenstein once remarked that 'Anyone who listens to a child's crying with understanding will know that psychic forces, terrible forces, sleep within it, different from anything commonly assumed. Profound rage and pain and lust for destruction' (1998, p. 4e). This may seem exaggerated; but we should bear in mind that our conception of infancy should allow for more than the notion of adorable babies we are all subject to. In particular it should also allow for the development of the astonishing aggression, hatred, and cruelty that we know to characterize our species, particularly as we engage in group conflict. It should therefore not surprise us if such emotions are also rooted in infancy. And researchers on aggression now seem agreed, as Tremblay reports, that aggression is at its most impulsive and forcible early in life, so that from infancy onwards 'rather than learning to physically aggress, children are learning not to physically aggress' (2004, p. 403).

In addition such early and survival-promoting expression of rage, fear, and distress at separation occur during the first postnatal stage of *parent/offspring conflict* (Trivers, 2002, discussed in relation to psychoanalysis in Hopkins, 2003, 2004), and hence when the infant's own genetic interests are most strongly opposed to those of its mother, father, and siblings. From the point of view of the infant and its genome, the mother's body and her will are the key to all resources. Their subjugation and exploitation will enable it to thrive, and without this it risks wasting and death. So it is not beyond possibility that at this time the infant should represent the mother's body as comparable to a territory it must conquer to live, and the father and other siblings (real or imaginary) as potentially life-threatening rivals, to be dealt with later.

## 20. Bayesian explanation in infantile experience

Finally, we must consider that it may well be deeply in the nature of the case that in early infancy the mother may be (proto-) hated and (proto-) blamed in her infant's mind or imagination, and very far in excess of her actual shortcomings or derelictions. For her infant's Bayesian brain *must perforce from the beginnings of consciousness seek to represent a cause for every experience of anxiety, suffering, and pain*. And what more salient candidate can there be, than some version of the breast and/or body of the mother the baby is already shaping its brain through learning to represent? Likewise we must consider that the infant may already be deploying early infantile versions of the high-level principles which will later govern the representation of faces and houses in the way we saw in the experiment above. In this case as the infant's experience oscillates between bad and good, its developing brain may at first construct *different* early episodic real and/or virtual objects as causes of its radically differing—some times very good, sometimes very bad—episodes of experience. (Early conceptual and emotional developments are discussed in more detail in Hopkins, 1987.)

Thus it should be regarded as a serious possibility on a Bayesian account that the infant might imagine a *very good* breast or maternal figure as the cause of its good or pleasurable experiences, and a distinct and *very bad* breast or maternal figure as cause of bad; and these would go with correlative experience of itself as in relation to such part-objects as well.[24] For in the case of the human infant, as our discussion from the next section onwards will indicate, we must consider not only its developing model of its mother or other carers, but also its developing image of itself and its own internal states.

These will clearly have complex interrelationships, but it seems likely that insofar as the infant feels itself as in contact with others who are good, it will more likely structure its own model of itself accordingly, and similarly for bad. (We will consider some evidence relating to this shortly.) Moreover insofar as the infant's (or child's) dominant model of itself excludes other models, we may expect the suppression of the excluded models to approximate Freudian repression very closely.

So—to take one of many possible scenarios—suppose an infant or child does form an image of its mother or father which provokes its own anger, resentment, and fear to a very high degree. And suppose also that the child needs to cooperate with that same parent, and also has love and affection for her or him, so that models in which child and parent apparently have good relations dominate the alternatives. In this case the child will have a genuine but repressed emotional conflict, in which feelings of anger, resentment, and fear—like input in the face/house case—will remain unconscious but permanently liable to activation in its mind. (Will remain, in Bayesian terms, likely to arise as ineradicable error-signal apparently contradicting the dominant model.)

## 21. Emotional conflict in infancy

We have good reason to believe that there are such conflicts. For we have just seen that during early infancy the infant directs powerful positive and negative emotions towards one and the same thing, namely its mother. This would seem to constitute a kind of natural liability in our species to

emotional conflict of this kind. So such conflicting emotions, and the representations which drive them, would seem to require to be resolved or mitigated by the time the infant comes to conceive of its mother as a single enduring object—for otherwise the infant would scarcely be able to relate to her in a coherent way.

Experiments on anger suggest that the baby's developing representations progressively regulate its emotions in this way. In particular, as the baby comes to organize a representation of its mother as bodily and psychologically whole, it changes the expression of anger from direction at bodily parts, so that by seven months it directs anger to the face of the person with whom it is angry; and it does this with a selectivity which shows that it has come to depend on the mother for comfort in coping with the intrusions of strangers, and so is liable to be particularly angry when she fails to play this role.[25] Also some evidence suggests that the baby begins to represent its mother as a single lasting (and therefore unique and irreplaceable) being during the fourth month of life, as Melanie Klein, the psychoanalyst who laid greatest emphasis on this development, hypothesized.[26]

## 22. Unresolved conflict and insecurity in attachment

Still the resolution achieved in early infancy is often strikingly incomplete, in the sense that representations laid down before the end of the first year may leave the individual liable to emotional conflicts which remain active throughout life. This is demonstrated by the basic measure of security of attachment, the 'strange situation' procedure devised by Ainsworth. This is used on infants of 12 months, so that its administration has been preceded by a series of typical developments. These include (1) the phases of regulation of anger apparent by seven months;[27] (2) those of distress at separation from the mother and fear of strangers which arise together at about eight months; and (3) the consequent consolidation of joint attention in an intersubjective and more fully communicative and cooperative relationship with the mother by about 10 months.

In the strange situation the mother cooperates with the experimenters in exposing the infant to successive short episodes of (1) encountering a stranger, (2) being left with the stranger, (3) being left entirely alone, and (4) being left entirely alone and then having to cope with the attentions of the stranger. So this procedure (in which each episode is terminated if it proves too upsetting) rouses the distress at separation and fear of strangers the infant has recently overcome, and with this its desires for comforting contact with the mother with whom it has recently consolidated a cooperative relationship. But of course it also rouses the anger the infant has long shown towards the mother whenever she defects from the protective and comforting roles on which the infant has come to rely, and left it, as on this occasion, alone, fearful, and at the mercy of a stranger in increasingly stressing ways.

The criteria demarcating secure from insecure attachment, in turn, mainly consist in expressions of conflict as between the anger and fear prompted by the procedure and the infant's desire to be comforted. Babies designated as secure resolve this conflict fairly readily despite their evident distress, and are soon comforted and able to return to exploration and play. Avoidant infants, by contrast, may seem unaffected by separation, but 'stiffen' with anger when mother tries to comfort, and consequently remain stressed for longer. Ambivalent infants alternate 'bids for contact with signs of angry rejection'; and disorganized infants seem 'incoherent', making 'interrupted

movements' or 'contradictory sequences or simultaneous behavioral displays' while giving 'indications of fear/apprehension' towards the mother. (Solomon and George, 2008, p. 387).

This indicates that the behaviours criterial for insecurity of attachment can also be seen as manifestation of early emotional conflict, rooted in images of the parents, and particularly the mother. These early patterns of conflict, in turn, can be seen to influence behaviour and development in myriad and often deleterious ways, and throughout the whole of life (Cassidy and Shaver, 2008, III, IV, V). Indeed the most serious cases of conflict, those exhibited by infants classed as disorganized, seem to exhibit a kind of oscillatory incoherence reminiscent of an internal version of the face/house example. Their contradictory sequences of behaviour, often seeming to attempt approach while manifesting fear and/or avoidance at the same time, seem just the sort of sequences which might flow from failing management of conflict in regard to *experience of emotion felt towards the mother*. (And for such infants more successful management of conflict, tellingly, seems to come only years later, and in the form of a permanent predisposition to behaviour which attempts to *control* the untrusted object of emotion, often by violent means.[28])

## 23. Internalization of relationships by the creation of imaginary internal figures (virtual others)

To understand the nature of the superego we must also consider another psychoanalytic claim. This is that we humans achieve our remarkable sociality partly by a particular use of the imagination. From early in life, and even when we are alone, we constantly imagine ourselves as in relation to others—virtual internal others—who have various kinds of relationships to us in our minds. We thus constantly in effect construct internal models of ourselves as in relation to others. We can use such models both for regulation and for learning; for in establishing such virtual figures, good and bad, we thereby create internal sources of reward and punishment, and hence of experience which can be evoked in a variety of simulatory ways.

It is easy to see this in the play of children. Thus take a child who watched her mother breaking eggs to make a cake, and was told not to break more eggs herself. She was later found saying 'No!' (as to an imaginary figure) and then turning and gleefully breaking an egg, and repeating the process again and again. In this we can see her as reworking and modifying the experience of moral prohibition, by enacting the role of prohibitor (in identification with the mother from whom she had received a prohibition shortly before) and then enacting the role of prohibitee, while replaying the situation as one in which the prohibitee obtains gratification by defying the prohibitor. (And of course there might be some significance in the fact that the prohibited objects were eggs, and so things which might be unconsciously imagined as potential siblings.)

In all this, the sensory inputs of the original episode of conflict between parent and child were being both internalized and modified. They were being reworked in terms of experiences, feelings, and actions on the part of the self as in relation to internal imaginary figures who were saying 'No!' to one another and also having 'No!' said to themselves. Freud (1920) discusses an earlier but similar example, concerned with the reworking and management of separation distress, in the game of an 18-month-old child; and the countless roles we can see children assume, repeat, and modify in

their imaginative play—as good mother or bad sister to a doll, as destroyer of an attacking monster, etc.—testify to the ubiquity of this phenomenon.

### 24. The importance of internalized punishment

Together with Klein (Freud, 1930, p. 130, 138; Klein, 1946) Freud took such virtual internal figures to be laid down from infancy in proto-conceptual memory from early and bodily phases. These figures were also able to produce *virtual sensory input*, as we observe in play and also regularly produce for ourselves when we daydream, ruminate, talk to ourselves, etc. (And we gain a Freudian—and perhaps depressing—perspective on the nature of such imaginative activity when we consider that aids and amplifications for imaginative engrossment in forms of sexuality and aggression are particularly popular on the internet.) Freud's and Klein's observations on the nature of such imagined figures in the minds of children also partly overlap with work in attachment.[29]

Freud described the 'good' figures laid down in this way in terms of an ego ideal. But he also found that the creation of *punitive, cruel, and moralistic figures* of this kind served as a principal means by which individuals regulated their aggressive impulses towards members of their families and other ingroups. Such figures were in effect internal repositories of the child's own aggression, as personified in images of others as potentially punitive and retaliatory. But by imagining itself as in relation to such figures, the child modified its own dispositions to aggression via fear of retaliation and punishment from dominant others, and by the development of guilt, shame, remorse, and other social emotions towards them.

In this case, however, psychological investigation showed the relevant internal figures to be extraordinarily dominating, punitive, and cruel. (And they often appear as monstrous and frightening in nightmares, such as that of the terrifying paternal figure described in Obama, 2008, 370ff.).

### 25. Internalized punishment in depression and schizophrenia

The internal ferocity of such self-directed aggression often appears clearly in depression and schizophrenia, and in both unconscious and conscious forms. Thus Elyn Saks (2008) describes her depression and schizophrenia in terms the internalization of moralistic aggression. As she became depressed, her thoughts started to run along lines such as *I am not sick. I'm just a bad, defective, and evil person. Maybe if I would talk less I wouldn't spread my evil around* (Saks, 2008, p. 58). They then went further, e.g. to *I am a piece of shit and I deserve to die. I am a piece of shit and I deserve to die. I am a piece of shit and I deserve to die* (Saks, 2008, p. 61).

### 26. Depression, self-directed anger, and the superego

That these expressions of self-dissatisfaction were also instances of aggression directed by her against her own self emerged particularly clearly, when with antidepressant medication her depression lifted for a time. She told her doctor 'Strangely, I feel less angry', and reports 'Not until that moment did I realize how much rage I had felt, directed mostly at myself' (Saks, 2008, p. 69).

This role of aggression was described clearly by Freud, where he says that the depressed individual 'represents his ego to us as worthless, incapable of any achievement and morally

despicable; he reproaches himself, vilifies himself and expects to be cast out and punished ... We see how in him one part of the ego [later to be called the superego] sets itself over against the other, judges it critically, and, as it were, takes it as its object' (1917, pp. 246-7).

### 27. Disintegration of the superego in schizophrenia

Saks' passage from depression into schizophrenia (or depressive psychosis) consisted partly in such a superego disintegrating into a group of virtual others who were insidious moralistic persecutors. Thus she describes how her internal presences began to multiply and change their role, as she herself began to lose her sense of agency in relating to them. As she says 'thoughts crashed into my mind like a fusillade of rocks someone or something was hurling at me—fierce, jagged, and uncontrollable ... *You are a piece of shit. You don't deserve to be around people. You are nothing. Other people will see this. They will hate you. They will hate you and want to hurt you. They are powerful. You are weak. You are nothing*' (Saks, 2008, p. 83).

Finally she 'began to feel I was receiving commands' from 'shapeless powerful beings that controlled me with thoughts (not voices) that had been placed in my head*. Walk through the tunnels and repent. Now lie down and don't move. You are evil* (Saks, 2008, p. 84). As she was so evil she was commanded to inflict pain on herself, and accordingly started burning herself in various ways, unable to tell others why. At last she spent most of her time alone 'in the music room or in the bathroom, burning my body, or moaning and rocking, holding myself as protection from unseen forces that might harm me' (Saks, 2008, p. 86).

### 28. Conflict and Freudian wishfulfilment

This brings us to Freud and *unconscious and internalized emotional conflict*, which Freud related to free energy.[30] Given the stage-setting so far, we will be able to address such conflict only briefly. We can start with a simple example closely related to the one we have already worked through.

### 29. Wishfulfilment and the management of conflict

Freud observed that during the night after he had eaten anchovies or some other salty food he was liable to dream *that he was drinking delicious cool water.* After several repetitions of this dream, he would wake up, feel his thirst, and get up to get a drink. This dream is a clear example of what Freud regarded as *wishfulfilment*: that is, as a representation

(i) caused by, and

(ii) representing the satisfaction of,

One or more of the agent's desires or wishes.

The desire in this case was Freud's desire to drink, which evidently caused him to wake after several repetitions of the wishfulfilling dream. The dream seems to have temporarily pacified this desire, which it also entirely masked from his dreaming consciousness, together with the thirst in which it originated.[31] Freud took this wishfulfilment as produced by his ego in order to manage a conflict between his thirst and his wish to sleep—or again between his thirst and the homeostatic mechanisms protecting sleep, to which he assimilated dreaming.

We can see some aspects of this conflict-managing process by contrasting the pattern of this dream with that of rational and successful action abstracted above. For such action we have

A des P [A drinks] → P [A drinks]→ A exps, bels P [A drinks] → A des P pacified.

In this, as we supposed above, the experiences of satisfaction predicted by the desire serve to pacify it, while its actual satisfaction obtained by drinking addressed the homeostatic imbalance in which it was rooted. In the dream, by contrast, we have

A des P [A drinks] → A *dream-exps bels* P [A drinks] → A des P *temporarily* pacified.

Here, in Freudian terms, the ego (= generative model in one role) apparently short-circuits the route which in action goes via real satisfaction, by producing an *illusory or hallucinatory version of the experience of satisfaction* predicted by the desire. This illusory experience of satisfaction, on Freud's account, permits sleep to continue.[32]

The dream thus instantiates an internal version of suppression/repression such as we saw in the house/face case and have elaborated in Freudian terms. In this case, however, the brain is dealing with an *internal* cause of sensory input which would be represented in consciousness as the experience of thirst or a desire to drink. So it has apparently repressed (and suspended the operation of) this desire rapidly and directly, *by producing an internal representation of the experience of satisfaction the desire predicts*. This, as Freud supposed, would seem tailored by his ego to enable him (for the time being) to sleep on; and it is done as the Bayesian brain would do, if, as seems possible, it was acting in the interests of homeostasis to keep motivational arousal from causing what it (as ego) calculated would be an uneconomic interference with sleep. But after a short time, apparently—thirst being such a demanding internal cause—the calculations changed in favour of satisfying the desire, and woke Freud up.

From the point of view of Freud's rational consciousness, however, this neurologically intelligible way of managing conflict related to internal sensory input appears as a kind of *perfect and all-encompassing miniature hallucination*, in which the deluded dreaming subject utterly obliterates both what is happening in his mind and how things are in the world. For if we take things in commonsense terms, the real underlying state of the dreamer's mind is that he is (unconsciously) thirsty and wanting a drink, and the relevant fact about the world is that he is lying supine in bed and doing absolutely nothing about this. At the same time, however, his dreaming brain (as ego) is producing a *double denial of reality*, in which he imagines that he is not thirsty but rather enjoying the slaking of thirst; and that he is not passive, frustrated, and asleep, but rather awake and experiencing his own activity in satisfying his desire. So overall the brain (generative model, conceptual system, ego) is temporarily producing a situation such that if that situation were prolonged it would die.

This double denial of reality is inherent in Freudian wishfulfilment. In dreams it is clearly harmless, and indeed one might be inclined to suppose that the intense wishfulfilling illusions of dreaming play a role in some form of learning, perhaps in coordination with the process of homeostatic synaptic 'renormalizing' which Greene and Frank (2010) consider in connection with slow-wave sleep. The situation, however, is otherwise in symptoms of mental disorder, as we can see in a slightly more complex case.

## 30. Symptom, id, and superego

The main symptom of Freud's (1909) patient the Rat Man was his compulsive involuntary repetition of episodes of vividly imagining—as if stuck in a waking nightmare—that his beloved (and long-deceased) father and/or the woman whose affection he sought were being subjected to a terrible torture, in which rats ate their way into his body from behind, causing an agonizing death. This, not surprisingly, made him anxious, guilty, and depressed; and he constantly sought to prevent or undo the occurrence of this torture via a variety of obsessional activities.

Since this symptom was a cause of guilt and depression, the wishfulfilment may not be as obvious as in the dream of drinking. Still it is quite within the scope of common sense, confronted with someone who compulsively represents another as undergoing a terrible torture, to suppose that hostility on the part of the former towards the latter may be somewhere in the offing. (Similarly one might expect to find sexual desire in the case of somebody who compulsively imagined sexual activity: and compare how we react to in real life when we find that a priest or pedagogue assists his own imaginings with pornography involving children.) And although Freud's analysis of this case is too complex to be discussed in detail here, he did encounter a broad range of evidence that the Rat Man harboured deep unconscious hostility towards his father, and that this was rooted in images of his father as frightening, punitive, and prohibitive, which he both remembered and projected on to Freud during the course of his analysis (Freud, 1909 as discussed in Hopkins, 1982).

On this account the symptom is analogous to the simple dream discussed just above. We have the pattern:

A des P [A's father tortured] → A *imaginarily* exps, bels P [A's father tortured] → A's des P *temporarily* pacified.

As in the example of the dream, this pacification of desire can be seen as a Bayesian repression serving to manage conflict. In this case, however, the conflict to which the pacified desire is part is a full-fledged Freudian conflict of the kind considered at the outset, involving basic emotions, long-term (but repressed) autobiographical memory, and a ferociously self-critical and thought-inhibiting part of the self. So here the conflicts among the superego, ego, and id, as sketched earlier in our proto-Freudian description of the face/house case, can be seen as the Freudian real thing.

## 31. The conflicting models in this case

In this example the dominant conceptual model—as expressed in the patient's partly faulty first-person authority—represented the patient and his father as having always been affectionate best friends. For this reason, according to the patient, it was unbelievable that he should harbour any hostility towards his father, who, among other things, had always treated him gently. His associations, however, sometimes qualified this: after one denial, for example, he recalled a story about a woman who had wished that her sister might die so that she could marry her husband, and had committed suicide for being so viscious. He said that it would be fair if he too were to die because of his imaginings, for he deserved nothing less.

As this indicated, there was an alternative and repressed model of his relationship with his father, which had been active from his early childhood. This emerged and was revised and partly dissipated in the course of his analysis. In this recessive model the father was represented as punitive, prohibitive, and frightening, and the child as his terrified victim.

**32. The ego and the id**

This conflict in the patient's feelings towards his father was reflected in one between the ego, here taken as the set of conceptual models dominant in the patient overall, and his id, taken as the locus of the subcortical sources of interoceptive input from homeostasis and emotion, particularly rage and fear)., These emotions were apparently mediated by conflicting models of himself as in relation to his parents, presumably formed in infancy and childhood, as reviewed above. For since his images of his father as punitive and terrifying were inconsistent with the dominant model, they—together with the feelings and desires for retaliation they aroused in the patient—had long been excluded from conscious awareness. In consequence they remained liable to activation in which, in dynamic conflict with the ego, they pressed upwards for expression in consciousness (as a signal of error which was unintelligible on the dominant model, and so could not be explained away).

## 33. Childhood Conflict.

The patient reported that he had been obsessive, depressed, and preoccupied with his father's death since the age of six, which he described as 'the beginning of my illness'; and this was apparently linked with representations of his father and his own sexual gratification as in some sort of lethal opposition. Thus he described how at six he wanted to see girls naked, but had 'an uncanny feeling' that if he thought such things something bad might happen, which, as in his present illness, he had to prevent—such as, that his father might die. Thus, as he said, 'Thoughts about my father's death occupied my mind from a very early age and for a long period of time, and greatly depressed me'. (1909a, p.162)

The model from which such thoughts were drawn at six had apparently remained active but repressed in later life as well, as illustrated by his thinking, while first having intercourse, that 'One might do anything for this—murder one's father for example'. (1909b, p.264). It seems to have been activated in the particular way that led to his breakdown by his hearing the 'cruel Captain' describe the Rat torture applied to prisoners of war. As he heard the account he imagined that the lady he venerated and his father were being tortured in the same way, felt that he had now urgently to prevent this (even though his father had been dead for many years), and began a series of obsessional acts aimed at doing so.

## 34. Thinking and the superego

Again as in the case with which we began, the patient (or his ego) was apparently capable of forming a concept, and engaging in a series of thoughts, which would explain his feelings and would serve to render them conscious. The first step, as Freud presented matters, was for the patient to consider that his imagining his father tortured in this distressing way might express hostility to his father which was in conflict with the love he also felt, and to try to explain this

situation. (This could be done via the hypothesis that the hostility had been precipitated in some forgotten era of childhood, before his preoccupation with his father's death began.)

The patient was able to consider this without difficulty in thinking of his lady. But he could not do so in the case of his father, even though he felt intense guilt towards him. This was an indirect indication of the way his inability to think about this topic (to mentalize, as discussed in the chapter by Fonagy and Luyten, this volume) was a consequence of fear and guilt generated by his superego.[33] So his denials continued even after he acknowledged that he regularly used his rat phantasy to attack people to whom he was hostile, including thinking when he first heard Freud's fee 'So many florins, so many rats'. (1909a, p. 213) Indeed the denials only stopped after the analysis brought forward material we can see as relating to his superego, although this case history was written long before Freud explicitly framed this concept.

In a particularly striking and dramatic episode, the patient came to feel terrified of Freud, feeling him to be a potential murderer, who might be about to 'fall on him like a beast of prey, to search out what was evil in him'. At the same time he began to remember and relive a beating he had received from his father as a little boy, when he had wet his parents' bed while lying between them. The reference to a beast which *searched out evil* by biting into the body enables us to see this as an image of the Rat Man's own (oral and bestial) superego, which was almost as murderous and sub-human as the rats he imagined attacking others. This is, of course, a different image than that of the terrifying figure in Obama's dream, or again the invasive persecutors whose presence Saks sometimes felt. But these figures illustrate a continuity between normal dreams, paranoid depressive phantasies, and the kind of phantasy experienced as real in psychoanalytic transference. (And when comparable transference phantasies were active in Saks, so that she was feeling her analyst as a potential murderer, she carried a knife to her sessions—which, of course, she never used.)

### 35. Revisions in conception and emotion

This was a turning point in the analysis, which apparently enabled the patient to revise his image of his father, and so to continue to love him while accepting that he had thought him terrifying and dangerous as a child, and had perhaps wanted to hurt him in consequence. Likewise it enabled him to modify the anxiety and guilt engendered by his superego, and so to think more freely in talking with Freud.

On this account, therefore, the same kind of conflict-engendering imagos as drove the wishes expressed in the patient's symptom had also been internalized to form a superego which punished him for his aggressiveness towards his father while at the same time as making it impossible for him to think about this aggression and so to understand it better. The effect of these imagos was thus to keep his ego oscillating between the imagined torture which pacified his uncorrected childhood rage and the guilt and depression he felt for imagining such things.[34] The cycle ended only when the imagos were re-experienced, reconceived, and so altered in the way they produced emotion, in his work with Freud.

### 36. Freudian wishfulfilment and pacificatory repression

Why? In an explicitly Bayesian context a further answer suggests itself, which coheres with accounts derived from Freud. We have already seen how emotional conflict involves, as well as free energy, a kind of situation we may suppose our generative models function to avoid.

This was exemplified above in the overtly contradictory behaviour apparently produced by conflicting internal models of self and other maintained by infants with disorganized attachment. Given that the conflicting desires managed by the Rat Man's brain were simultaneously expressed in his imaging his *deeply loved father* (true in the long dominant conceptual model, and also true on realistic reflection over the course of his life) father being *repeatedly subjected to terrible torture* (expressing rage truly felt in early life, as registered in repressed and consequently active but recessive models) we can envisage that the expression of such desires via the patient's motor system would have been incoherent.

So we may perhaps be able to see this patient's brain (conceptual system, generative model, ego) as pacifying these desires as soon they as they arose by the most direct means possible, that is—and as in the simple dream we considered above—by falsely but immediately representing the predictions to which the repressed desires gave rise as having been fulfilled. In this way the brain succeeded in suspending the working of such desires, in the absence of any real attempt at satisfaction. This dreamlike process of pacification, however, was also the symptom which rendered the patient anxious, depressed, and obsessional. So here, on this Bayesian account, the mechanism of expression/suppression/repression by which the brain pacifies desire in such a conflict would also partly constitute the illness from which the patient suffers.

Given the space available, this sketch lacks detail. Still it may serve as an illustration of principle. Also, as consideration of hysteria and hypnosis suggest, a similar account might be applied to many other phenomena, including the case of 'Blind and sighted in the same person', which initially attracted the attention of Berlin and Koch. So in seeking to apply the kind of account illustrated here to further cases we might start to understand the Freudian unconscious as the natural product, in our conflicted species, of the management of conflict by the Bayesian brain.

References

Aydede, M. (2010). The language of thought hypothesis. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2010 edn. Online: http://plato.stanford.edu/archives/fall2010/entries/language-thought/.

Beebe, B., Jaffe, J., Markese, S., Buck, K., Chen, H., Cohen, P., Bahrick, L., Andrews, H., and Feldstein, S. (2010). The origins of 12-month attachment: A microanalysis of 4-month mother-infant interaction. *Attachment and Human Development*, 12(1), 3–141.

Berlin, H and Koch, C. (2009). Neuroscience meets psychoanalysis. *Scientific American Mind*, April/May, 16–19.

Bower, T. (1977). *Development in Infancy*. San Francisco, CA: W.H. Freeman.

Braten (1998) (ed.) *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge: Cambridge University Press

Campos, J.,Barret, K., Lamb, M., Goldsmith, H., and Stenberg, C. (1983) Socioemotional Development. In P. Mussen (ed) *Handbook of Child Psychology*, vol 3, New York: John Wiley.

Carhart-Harris, R. and Friston, K. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283.

Carhart-Harris, R., Mayberg, H.,, Malizia, A., and Nutt, D. (2008). Mourning and melancholia revisited: correspondences between principles of Freudian metapsychology and empirical findings in neuropsychiatry. *Annals of General Psychiatry*, 7, 9.

Cassidy, J. and Shaver, P. (2008). *Handbook of Attachment*. London: Guildford Press.

Clark, A. Whatever Next? Unpublished Manuscript, communicated 2011.

Craig, A. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, 10, 59–70.

Craig, A. (2010). The sentient self. In special issue, *Brain Structure and Function*, 214, 563.

Damasio, A. (1999). *The Feeling of What Happens*. London: Vintage Books.

Damasio, A., Grabowski, T., Bechara, A., Damasio, H., Ponto, L., Parvizi, J, and Hichwa, R. (2000). Activity during the feeling of self-generated emotions. *Nature Neuroscience*, 3(10), 1049.

Dayan, P., Hinton, G., Neal, R., and Zemel, R. (1995). The Helmholtz machine. *Neural Computation*, 7, 1022–1037.

Dennett, D. (1991). *Consciousness Explained*. New York, NY: Little Brown Publishers.

Downes, S. (2010). Evolutionary psychology. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2010 edn. Online: http://plato.stanford.edu/archives/fall2010/entries/evolutionary-psychology/.

Doya, K., Ishi, S., Pouget, A., and Rao, R. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Boston, MA: MIT Press.

Freud, S. (1895/1957). Project for a Scientific Psychology. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, *Volume I*, p. 297. London: Hogarth Press.

Freud, S. (1900/1957). Analysis of a specimen dream. In J. Strachey (ed.) *The Interpretation of Dreams,* in Strachey, J., ed., *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, *Volume V*, pp. 96–121. London: Hogarth Press.

Freud, S. (1909a/1957). Notes upon a case of obsessional neurosis. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, *Volume XXII*, p. 54. London: Hogarth Press.

Freud, S. (1909b/1957 'Original record'). Notes upon a case of obsessional neurosis: Addendum, original record of the case. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, *Volume X*. London: Hogarth Press.

Freud. S. (1917/1957). Mourning and melancholia. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: Hogarth Press.

Freud, S. (1920/1957). Beyond the pleasure principle. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: Hogarth Press.

Freud, S. (1930/1957). Civilization and its discontents. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: Hogarth Press.

Freud, S. (1933/1957). New introductory lectures on psycho-analysis. In J. Strachey (ed.) *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, *Volume XXII*. London: Hogarth Press.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9), 1325–1352.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society*, 360, 815–836.

Friston, K. (2010a). The free energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, 11, 127–138.

Friston, K. (2010b) Embodied inference: or 'I think, therefore I am, if I am what I think'. In W. Tschacher and C. Bergomi (eds) *The Implications of Embodiment (Cognition and Communication)*. New York, NY: Imprint Academic.

Friston, K. and Stephan, K. (2007). Free-energy and the brain. *Synthese*, 159, 417–458.

Friston, K., Daunizeau, J., Kilner, J., and Kiebel, S. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3), 227–260.

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104, 137–160.

Fuster, J. (2009). Cortex and memory: the emergence of new paradigm. *Journal of Cognitive Neuroscience*, 21(11), 2047-2072.

Gertler, B. (2011). Self-knowledge. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Spring 2011 edn. Online: http://plato.stanford.edu/archives/spr2011/entries/self-knowledge/.

Godfrey-Smith, P. and Sterelny, K. (2008). Biological information. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2008 edn. Online: http://plato.stanford.edu/archives/fall2008/entries/information-biological/.

Greene, R. and Frank, M. (2010). Slow wave activity during sleep: functional and therapeutic implications. *The Neuroscientist*, 16(6), 618–633.

Hinton, G., Dayan, P., Frey, B.J., and Neal, R. (1995). *The Wake-Sleep Algorithm for Unsupervised Neural Networks Science*, 268, 1158–1160.

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, 108(3), 687–701.

Hopkins, J. (1982/1995/2007) Introduction: philosophy and psychoanalysis. In R. Wollheim and J. Hopkins (eds) *Philosophical Essays on Freud*. Cambridge: Cambridge University Press. Partly reprinted in C. MacDonald and G. MacDonald (eds) (1995) *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford: Basil Blackwell.

Hopkins, J. (1987). Synthesis in the imagination: psychoanalysis, infantile experience, and the concept of an object. In J. Russell (ed.) *Philosophical Perspectives on Developmental Psychology*. Oxford: Basil Blackwell.

Hopkins, J. (2000a). Psychoanalysis, metaphor, and the concept of mind. In M. Levine (ed.) *The Analytic Freud*. London: Routledge.

Hopkins, J. (2000b). Evolution, consciousness, and the internality of the mind. In P. Carruthers and P. Chamberlain (eds) *Evolution and the Human Mind: Modularity, Language, and Meta-Cognition*, pp. 276–298. Cambridge: Cambridge University Press.

Hopkins, J. (2003). Evolution, emotion, and conflict. In M. Chung (ed.) *Psychoanalytic Knowledge*. London: Macmillan/Palgrave Press.

Hopkins, J. (2004). Conscience and conflict: Darwin, Freud, and the origins of human aggression. In D. Evans and P. Cruse (eds) *Emotion, Evolution, and Rationality*. Oxford: Oxford University Press.

Hopkins, J. (2007). The problem of consciousness and the innerness of the mind. In M. McCabe and M. Textor (eds) *Perspectives on Perception*, pp. 19–46. Frankfurt: Lancaster Publishers.

Horgan, T. and Teinson, J. (1999). Rules and representations. In R. Wilson and F. Kiel (eds) *The MIT Encyclopedia of the Cognitive Sciences*, pp. 724–726. Cambridge: MIT Press.

Kant, I. (1781, 1963). *The Critique of Pure Reason*. (N. Kemp-Smith, Trans.). London: Macmillan.

Kernberg, O. (2009). An integrated theory of depression. *Neuropsychoanalysis*, 2(1), 76–80.

Klein, M. (1946, 1997). Notes on some schizoid mechanisms. In *Envy and Gratitude and Other Works,* The Melanie Klein Trust, London: Vintage.

Klein, M. (1997) Love, Guilt, and Reparation. The Melanie Klein Trust, London:Vintage.

Klein, M. (1998) The Psychoanalysis of Children. The Melanie Klein Trust, London: Vintage.

Klein, M. (1998) Narrative of a Child Analysis. The Melanie Klein Trust, London: Vintage.

Laurence, S. and Margolis, E. (2011). Concepts. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Summer 2011 edn. Online: http://plato.stanford.edu/archives/sum2011/entries/concepts/.

Libet, B. (1982). Subjective antedating of a sensory experience and mind-brain theories. *Journal of Theoretical Biology*, 114, 563–570.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529–566.

McLaughlin, B. and Bennett, K. (2010). Supervenience. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Summer 2010 edn. Online: http://plato.stanford.edu/archives/sum2010/entries/supervenience/.

Millikan, R. (1982). *Language Thought and Other Biological Categories*. Cambridge: MIT Press.

Millikan, R. (2005). The language-thought partnership: a bird's eye view. In R. Millikan (ed.) *Language: A Biological Model*, pp. 92–105. Oxford: Clarendon Press.

Obama, B. (2008). *Dreams from My Father (A Story of Race and Inheritance)*. Edinburgh: Canongate Books.

Pace-Shott, E., Solms, M., Blagrove, E., and Hanard, S. (2003). *Sleep and Dreaming: Scientific Advances and Reconsiderations*. Cambridge: Cambridge University Press.

Panksepp, J. (1998) Affective Neuroscience: The Foundations of Human and Animal Emotions Oxford: Oxford University Press.

Parvizi, J. and Damasio, A. (2001). Consciousness and the brainstem. *Cognition*, 79, 135–159.

Parvizi, J. and Damasio, A. (2003). Neuroanatomical correlates of brainstem coma. *Brain*, 126, 1524–1536.

Pitt, D. (2008). Mental representation. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2008 edn. Online: http://plato.stanford.edu/archives/fall2008/entries/mental-representation/.

Prior, V. and Glaser, D. (2006). *Understanding Attachment and Attachment Disorders*. London: Kingsley Publishers.

Robbins, P. (2010). Modularity of mind. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Summer 2010 edn. Online: http://plato.stanford.edu/archives/sum2010/entries/modularity-mind/.

Saks, E. (2008). *The Center Cannot Hold: My Journey Through Madness*. New York, NY: Hyperion.

Samuels, R. (2006). Is the mind massively modular? In R. Stainton (ed.) *Contemporary Debates in Cognitive Science*. Oxford: Blackwell.

Segal, H. (1978). *Introduction to the Work of Melanie Klein*. London: Hogarth Press.

Segal, H. (1981a). *Klein*. Glasgow: Fontana.

Segal, H. (1981b). *The Work of Hanna Segal*. New York, NY: Jason Aaronson.

Shea, N. (2007). Representation in the genome, and other inheritance systems. *Biology and Philosophy*, 22, 313–331.

Sherrington, C. (1906). *The Integrative Action of the Nervous System*. New Haven, NJ: Yale University Press.

Schore (2001) Effects of a secure attachment relationship on right brain development, affect regulation, and infant mental health. *Infant Mental Health Journal*, 22(1–2), 7–66.

Siewert, C. (2008). Consciousness and intentionality. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2008 edn. Online: http://plato.stanford.edu/archives/fall2008/entries/consciousness-intentionality/.

Smith, K.,, Berridge, K., and Aldridge, W. (2011) Disentangling Pleasure from incentive salience and learning signals in brain reward circuitry. PNAS July 5;108 (27):E255-64.

Solms, M. (1997). *The Neuropsychology of Dreaming*. Mahwah, New Jersey: Lawrence Erlbaum.

Solms, M. and Turnbull, O. (2002). *The Brain and the Inner World*. London: Karnac Press.

Solomon, J. and George, C. (1999). *Attachment Disorganization*. New York, NY: Guilford.

Solomon, J. and George, C. (2008). The measurement of attachment security and related constructs in infancy and early childhood. In J. Cassidy and P. Shaver (eds) *Handbook of Attachment*, pp. 383–410. London: Guildford Press.

Stoljar, D. (2009). Physicalism. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2009 edn. Online: http://plato.stanford.edu/archives/fall2009/entries/physicalism/.

Thagard, P. (2010). Cognitive science. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Summer 2010 edn. Online: http://plato.stanford.edu/archives/sum2010/entries/cognitive-science/.

Tremblay, R. (2004). The development of physical aggression in infancy. *Infant Mental Health Journal*, 25(5), 399–407.

Trivers, R. (2002). *Natural Selection and Social Theory*. Oxford: Oxford University Press.

Tseng, Y.W., Diedrichsen, J., Krakauer, J.W., Shadmehr, R., and Bastian, A.J. (2007). Sensory prediction errors drive cerebellum-dependent adaptation of reaching. *Journal of Neurophysiology*, 98, 54–62.

Waldvogel, B., Ullrich, A., and Strasburger, H. (2007). Blind und sehend in einer Person. *Nevenzart*, 78, 1303–1309.

Watt, D. and Panksepp, J. (2009). Depression: an evolutionarily conserved mechanism? *Neuropsychoanalysis*, 2(1), 93.

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wellman, H. (1990). *The Child's Theory of Mind (Learning, Development, and Conceptual Change)*. Cambridge, MA: MIT Press.

Wittgenstein, L. (1998). *Culture[35] and Value*. Oxford: Blackwell.

[1] I am particularly grateful to Geoff Hinton, Andy Clark, Jonathan Lear, and Aikaterini Fotopoulou, for providing points of discussion and criticism without which this paper would not have been written; to Tamas Pataki and Sam Guttenplan for penetrating and helpful criticisms of early drafts; and to Karl Friston for reading a late draft and generously expressing an encouraging degree of sympathy with the underlying argument.

[2] The simple structure of wishfulfilment, as later described in the text by reference to the dream of drinking, was also used by Freud (1900) to interpret his dream of Irma's injection. There, as discussed in Hopkins (1996, 1999a) this structure was used to integrate the series of memories which appeared in his free associations, and which also cohered in their origin from the deepest sources of his own shame and guilt. (As Freud said in his associations, he seemed to be 'collecting' such examples to 'bring against myself' in the dream.)

While the dream can indeed be understood as a straightforward wishfulfilment on the model of the dream of drinking, the explanation of the data Freud provides becomes more detailed and cogent once the wishfulfilment in the dream is seen as an instance of the mechanism of *projection,* which the dream clearly displays in representing the kinds of dereliction about which Freud was most guilty as in Otto. Likewise, and considering further theoretical developments, the dream seems a clear instance of Kleinian projective identification, as indicated by the way the bodies of figures who appear in the associations bear marks of physical invasion (the necrosis of the skin in the patient whose nasal swellings Freud had treated with cocaine; Irma's infiltrated skin as invaded by Otto's toxic sexual injection, Freud's own shoulder which ached in identification, etc.)

Finally the same concept appears in the wishfulfilling aspect of what Freud calls the primary process, by which the brain initially meets present needs with memories of past *experiences of satisfaction.* These experiences, which are particularly important in both Freudian and Bayesian accounts, provide one of the main links between them.

[3] If we understand mental states and processes in this way, we can also give an account of our *that P* mode of describing them. In this we are using representations to which we have public perceptual access (the words and sentences of our natural language) to describe the internal representations (beliefs, desires, etc.) which prompt and guide our behaviour, but to which we have no such access. (As, indeed, we do not normally have perceptual access to the internal workings of the brain or nervous system.) Commonsense psychology (theory of mind, etc.) thus employs sentences as audible or visible linguistic representations, in order to describe mental representations which govern our behaviour, but are inside us and so imperceptible to these senses. And given some antecedent capacity for verbal expression, it is easy to see how evolution might have nudged this towards the mind-articulating capacity we now enjoy (Hopkins, 2000b).

[4] It is perhaps worth stressing that there seems no reason to hold that in its deeper workings the brain employs anything like the 'language of thought' stressed in one tradition in philosophy and cognitive science (Aydede, 2010). For neural representation seems a matter of massive coordinated but differing forms of cellular and subcellular chemical and physiological activity, described by probabilistic functions. Thus as Andy Clark writes in his illuminating 'Whatever Next' (unpublished):

> Instead of simply representing 'CAT ON MAT' the probabilistic Bayesian brain will encode a conditional probability density function, reflecting the relative probability of this state of affairs (and any somewhat-supported alternatives) given the available information. This information-base will include both the bottom up driving influences from multiple sensory channels, and top-down context-fixing information of various kinds.

As I take Clark's description to imply, such probability functions would map to highly distributed physiological processes produced and used by very many hierarchically arranged networks of neural cells, which in turn would map to the very many things and situations they represent in many different ways at once, for the producing and using networks to discharge the functions for which the genes involved in their construction have been selected and maintained. If this is the basic representational situation in the brain there can be no reason to assimilate it to the use of sentence-like representations in a digital compiler.

Again the person-level hierarchies here do not seem to be *modular* (Robbins 2010) in the sense imagined in popularized evolutionary psychology (Downes, 2010). Rather as Fuster argues, we should take seriously the fact that modular accounts of cognitive functions are based on a 'definition of a module' which as regards memory and working memory 'is theoretically and empirically inconsistent with the recent literature' (2009, p. 2049). So apart from basic sensory and motor systems we should expect *(not a modular but) an hierarchically integrated processing organization*: as in the 'massive scaffolding' cited in the text). The notion of modularity has also been subject to serious philosophical criticism, as in Samuels (2006).

[5] There is often said to be an unbridgeable gap, as between the physical working of the brain and our conscious experience of ourselves in the world. As noted below, I take this supposed gap to involve an illusion produced by the working of the brain. But at this point one can say that insofar as there seems to be a gap, the Helmoltz/Bayes account partly consists in the claim that the brain itself crosses it, by transforming sensory input into conscious experience. This yields what seems a non-reductive form of supervenience physicalism (Stoljar, 2009), McLaughlin and Bennett (2010). This account plausibly supports token event-identity but not type-type identity. for as the quotation from Clark (unpublished) in the previous footnote suggests, the realization of types would appear so local both to environmental circumstance and to variation among brains that strict identities between non-ideopathic types would be ruled out, as argued in Hopkins (2007).

[6] This omits a number of complexities in the account by which Friston models this process.

[7] Fuster independently stresses the role of the side-by-side and backwards connections, urging, for example, that 'reentry is an integral part of the most plausible computational models of working memory' (2009, p. 2056).

[8] This description is of course very rough and covers a number of different approaches. In Friston's formulations representational optimization and error correction are done in accord with the principle of minimizing free-energy as a measure of surprise, which Carhart-Harris and Friston (2010) relate to Freud's discussions of bound and free energy. Such an account, as Friston (2010a,b) stresses, assigns a particularly encompassing role to Bayesian prior expections, and yields an understanding of the role of attention, dopamine, and evaluation or reward which contrasts with many 'reward maximizing' approaches. Likewise the conception of mirror neurons advocated by Friston et al. (2011) differs in many particulars from other versions. Nonetheless surprise-minimizing (= prediction error) and reward-maximizing accounts can be seen as falling within the broader Helmoltz/Bayes tradition.

[9] Thus relatively high-level processing involving conceptual metaphor seems to influence the brain's representation of the mind as a kind of internal but non-physical space located within a physical container (perhaps originating in interoceptive feedback from the skin, as suggested in Hopkins 2000a); and the same would hold for the overlap between the fields of conceptual metaphor and symbolism in psychoanalysis more generally.

Again, the 'multiple drafts' of Dennett's (1991) will be on file and constantly engaging in mutual revisions in many different levels in many different neural hierarchies; and these, as Dennet claimed, are to be understood as producing, and in that sense explaining, the whole of conscious experience.

[10] As argued in Hopkins (2007), the classical philosophical problem of consciousness arises from the apparent contradiction produced by our sense of experience as inner, phenomenal, subjective, and private to ourselves, as opposed to its distal objects, which are outer, physical, objective, and publicly available to all. The present account provides for the resolution of this problem by explicating consciousness in terms of the brain's image of itself, which provides our own from-inside images of our selves. (Hopkins 2000b). For the representation by each individual's brain of its own neural input naturally appears *to that individual* as phenomenal, as well as inner, subjective, and private, which this representation actually is; whereas the external objects presented in the representation are shown as outer, physical, and public, as in they fact are.

[11] As with much else in this essay I owe this example to Clark's (unpublished) 'Whatever next?'.

[12] There is often an ambiguity in Bayesian formulations, as to whether the brain is predicting the course of experience (predicting its own sensations) or predicting neural input to its own 'sensorium' by representing that input as sensations and other experiences caused in particular ways, for example by objects such as faces and houses. In fact we should take the brain as doing both, because in representing input as experience of any kind, it perforce also predicts both input and experience. Since both points hold we will ignore this ambiguity in what follows.

[13] In representing the input as conscious experience of a particular object the dominant model is said to explain it away, in the sense that aspects of the input predicted by the conscious representation are suppressed at lower levels in the hierarchy, while alternative conceptual explanations are inhibited at the higher. Insofar as the conscious experience is veridical and accurate in its predictions, the suppression it effects leaves no active residual. Unpredicted input, by contrast, is not silenced in the same way: it continues to be sent forward as error signal, and in that sense continues to press upward for conscious expression.

[14] Thus for example our awareness of our choices (say as causes of our very experiences of choosing) is synthesized only after the choices themselves. This is to be expected, since an event of conscious awareness of x involves the application of concepts to x, and so must in general occur (at least very slightly) after the *x* in question has itself occurred. This seems to have caused widespread puzzlement (Libet, 1982, 1985; Wegner, 2002).

[15] This is why each of us appears to him- or herself as the kind of self repeatedly postulated in philosophy: for example as the subject of Descartes 'I', or again Kant's transcendental self that synthesizes the manifold of sensible intuition. Such representations of the self reflect the way the brain represents perceptual input as experience of the self in the world.

[16] This suggests that while we may reasonably think of the ego as realized by the brain operating in default mode, as suggested by Carhart-Harris et al. (2008) and again by Carhart-Harris and Friston (2010), we should think of the real operative factor in the ego as the agent's conceptual system, as embodied in the underlying generative model. It seems to be this—and with it the emotions and thoughts that it serves to regulate—that is, at a flexible equilibrium in what is regarded as the default mode.

[17] The claim of conservatism here refers to the fact that the new representation is rejected because it does not fit with Bayesian assignments of prior probability over possible representations or concepts. These are presumably made on the basis of past experience, or are built in, e.g. as innate biases structuring neural processing. The Freudian superego may have a similar and innate structure; for it seems to realize an evolutionarily established direction of moral aggression against the self that may have evolved (together with a related direction of moral aggression against outgroups) by facilitating ingroup cooperation, as discussed in Hopkins (2003, 2004)

[18] And of course this case, and the present discussion, admit comparison with similar phenomena in the essays by Bazan and Snodgrass, by Oakley, and by Raz and Wolfson in this volume.

[19] Friston, Daunizeau, Kilner and Kiebel (2010) suggest that their model has radical consequences for the notion of action. On their account 'the central nervous system is not divided into motor and sensory systems but is one perceptual inference machine that provides predictions of optimal action, in terms of its expected consequences'. Moreover 'the only thing that action can affect is the prediction error at the sensory level. This means action can only suppress the weighted sensory prediction error variance' so that 'action is just there to explain away unexpected sensory prediction errors.' This, they hold, 'means we can replace the notion of *desired* movements with *expected* movements and understand action in terms of perceptual expectations.' But as we have seen, evolution has already built the required notion of expectation into the notion of desire, via our practice of describing desires in terms of the effects they are predicted to produce if acted on. So in 'explaining away unexpected prediction errors' actions satisfy desires by causing the experiences of satisfaction they predict, and thereby minimize the homeostatic or emotional disequilibria (= sources of free energy) which are their source. This is how 'ensuring our predictions become a self-fulfilling prophecy' keeps us in the attractors which avoid internally generated homeostatic surprise – which, as the case of thirst illustrates, is no surprise, in commonsense terms, to those that suffer it. For an example in which sensory prediction errors apparently serve as motor commands see Tseng, Diedrichsen, Krakauer, Shadmehr, and Bastian (2007).

[20] For the case of interoceptive input (1) above seems the experience by which sensory input is represented as a cause, and in this sense the initial analogue for the internal case of the Bayesian 'explaining away' of exteroceptive input stressed by Friston. So from the time the agent *experiences* drinking, and so pacifies the desire in (1) the 'free energy' initially put to work in the desire to drink can be said to remain bound while the underlying equilibration in (2) is effected. But it is also in the nature of such input that the desire suspended in the period between (1) and (2) should be subject to revival and/or strengthening, should (2) fail to occur—as in the psychoanalytic cases we will discuss later in the chapter.

[21] The link with the notion of explanation which makes speaking of explaining away appropriate in the exteroceptive case is partly retained here, for both the desire and its underlying homeostatic cause are ultimately pacified via the truth of predictions made by the brain in relation to them.

[22] Thus this account also coheres with broad outlines of theories of emotion and consciousness advanced by both Damasio and Panksepp, as described in the discussions of emotion, consciousness, and the self in Solms and Turnbull (2002). These accounts have recently been supplemented by work by Craig on interoception (2009, 2010).

[23] For early accounts of infancy highlighting maternal investment in cortical development see Schore (2001); and the essays by Trevarthan and others in Braten (1998) .

[24] So this might well be the origin of the origin of what Melanie Klein describes in terms of the splitting of the breast, and later the mother, into bad and good versions, as described and referenced in Segal (1978) and discussed in Hopkins (1987). Also if the infant made use of metaphorical representation as considered in Hopkins (2000a), such thinking might appear in metaphors of the mind as a container, as does the Kleinian notion of projective identification (Segal, 1978). Even the extremities of Klein's account of the baby imagining invading the mother's body to attack versions of the father and siblings within might be consilient with a combination of Bayesian representation and parent-offspring conflict as briefly sketched in Hopkins (2003, 2004).

In this context consider the Rat Man's phantasy (for which he expected retribution) of Freud's mother dead, with her breast impaled by the Rat Man's Japanese swords representing *marriage* and *copulation*, and Freud and his children eating away at the lower parts of her body, especially her genitals, like the rats of his own phantasy about his father's ongoing torture (Freud, 1909b, 'Original Record', p. 282). This is the kind of phantasied invasion of the mother's body later emphasized by Klein, and there can be no question of it having been produced in the Rat Man by Freud's suggestion.

[25] Thus consider some examples from Campos, Barret, Lamb, Goldsmith, and Stenberg (1983) : When someone makes a four-month-old baby angry by impeding its movements, the baby directs its rage *at the impeding hand*. So despite its impressive capacity for other-directed rage and fear, the four-month-old baby

seems not yet to have come to represent another's hand as part of, and so as animated by, an anatomically whole person (and is also an instance of the psychoanalytic notion of an emotional relation to a part-object, which should still apply at this age to the mother generally, and would particularly include her breast.) A seven-month-old baby, by contrast, directs its anger to the impeding agent's face. By this age, it seems, the baby has attained a more coherent representation of the human body, and one which enables it to relate emotionally person to person and face to face. And although the seven-month-old baby protests at being impeded by either its mother or a stranger, it is particularly upset when the mother impedes it after a stranger has done so. So by this time its anger is also regulated by its representation of its mother as providing, and itself as requiring, protection and comfort where strangers are concerned.

[26] Hopkins (1987) describes how these developments relate to theories held by Klein and Piaget. But as discussed there and also and briefly in Hopkins (2003, 2004) one experiment seems particularly relevant. Bower (1977, p. 217) describes

> A simple optical arrangement that allows one to present infants with multiple images of a single object ... If one presents the infant with multiple images of its mother—say three 'mothers'—the infant of less than five months is not disturbed at all but will in fact interact with all three 'mothers' in turn. If the setup provides one mother and two strangers, the infant will preferentially interact with its mother and still show no signs of disturbance. However, past the age of 5 months (after the co-ordination of place and movement) the sight of three 'mothers' becomes very disturbing to the infant. At this same age a setup of one mother and two strangers has no effect. I would contend that this in facts shows that the young infant (less than five months old) thinks it has a multiplicity of mothers, whereas the older infant knows it has only one.

These experiments do seem to admit interpretation as evidence that while at four months the infant takes its mother as a psychological other to whom it relates, it does not yet regard her as a single enduring person, as opposed to a potential multiplicity of presences whose spatiotemporal dimensions are as yet indeterminate. By five months, however, the baby apparently opposes uniqueness to episodic multiplicity, and starts to represent the mother (and by implication/identification its own self) as individual, continuous, and lasting.

If this is correct, then the four- to five-month consolidation of the mother's image via the concept of spatiotemporal numerical identity represents a synthesis in the imagination by which the baby integrates the major parameters of its internal and external worlds. We should regard this as a momentous event, particularly in light of the considerations about motivational conflict advanced here. As such it deserves fuller experimental investigation.

[27] Cf. the pattern of arousal of anger in relation to provocation by strangers at seven months in the previous note.

[28] For more on disorganized attachment, see Solomon and George (1999).

[29] Klein's collected writings appear in the bibliography with Klein (1946). For an introduction to her work see Segal (1978), Segal (1981a), or the single essay 'Melanie Klein's technique of child analysis' in Segal (1981b). As noted Hopkins (1987) contains discussion of Klein's ideas which relate to the argument of this paper. For work in attachment which can be related to some of the same emotional themes see 'Assessments of attachment based on the child's internal working model/representation' at pp. 109ff of Prior and Glaser (2006).

[30] The role Freud assigned free energy has enabled Carhart-Harris and Friston (2010) to relate the information-theoretic version of this notion to Freud's uses, so as to yield a field of evidence consilient with Freudian claims. But they omit to consider the role of conflict as a generator of free energy (prediction error) even though the role of conflict in neurosis and psychosis is widely acknowledged (Kernberg, 2009), and would fit with the data they survey. Likewise in their admirable 'Mourning and melancholia revisited' Carhart-Harris, Mayberg, Malizia, and Nutt (2008) seem to scant the role of the split-off part of the ego which was to become the superego as an internal source personal-level conflict ('self-reproaches and self-revilings') within the self, as illustrated by the material from Saks. Rather they stress only the (also very relevant) role of repression and object-loss instead. Thus they observe that one of the depressed patients who recovered from treatment-resistant depression almost instantly upon receiving stimulation in Cg 25 reported that the experience was like release from being 'locked in a room with 10 screaming children: constant noise, no escape'. This might well be taken to suggest that activity in Cg 25 also relates to the representation of painful internalized emotional conflict, of the kind to be discerned in relation to the superego, and also perhaps to parent/offspring conflict. But the datum was explained in terms of release of repression (of what?) instead.

[31] Freud did not describe wishfulfilment in terms of pacification; but it is clear that he regarded the fictitious *experience of satisfaction* as having this role. He introduced the notion as explaining his own dream of Irma's injection (1900, pp. 96–121), and as 'the first member of a class of abnormal psychical phenomena' including 'hysterical phobias, obsessions and delusions'. As is often the case with such advances, his paradigmatic expositions introduce data which in retrospect we can see as better explained by succeeding theories into which his original ideas were incorporated. Thus for example the wishfulfilment analysed in the Irma dream seems clearly also to be *defensive* (and against internalized conflict) and *projective* as well. Much of the material, for example, involves instances of 'lack of medical conscientiousness', mainly involving deadly or

harmful injections associated with his own activities, which Freud observes he 'seemed to be collecting to bring up against myself' in the dream. This 'collection' compared his medical derelictions to murders which might prompt talionic revenge, in the form of his own daughter's death (see 'this Mathilde for that Mathilde, an eye for an eye and a tooth for a tooth' in the associations to *I at once called in Dr M*.) Later Freud would have regarded this as the work of his own morally punitive superego; and by the end of the dream he had managed to identify himself with this superego, so as to declare 'one does not make injections of that kind so thoughtlessly: and probably the syringe [with which he dreamt Otto had injected Irma] was not clean.' So the wishfulfilment in Freud's paradigmatic dream can also now be seen as a defence against his own superego; and it was straightforwardly projective, since everything related to his own lack of medical conscientiousness (and worse) had been projected into Otto, whose remarks about Irma had roused Freud's guilt and prompted the dream.

The signs of damaging physical intrusion which go with this projection also mark it as a complex instance of unconscious Kleinian *projective identification* (Segal 1978) used to deflect depressive anxiety.

[32] The claim that dreams generally have this function, in relation to emotional/motivational arousal which occurs regularly in sleep, has recently been pursued by Mark Solms and others (Pace-Shott et al., 2003; Solms 1997; Solms and Turnbull, 2002).

[33] The link between the persecuting internal figures constituting the superego and the capacity to think was stressed in the work of Bion (e.g. 1967). The claim here is that on a Bayesian model these 'earliest parental imagos' may also constitute assignments of prior probabilities which make certain kinds of thinking impossible, as seen in the face/house example.

[34] Feud's notes, far in advance of his theories at the time, provide evidence that the origin of the Rat Man's conflicts was to be found in his infantile imagos of his mother. For the episode in which he remembered his father as a fearful punisher, and experienced Freud in the transference as a murderous moralistic invasive beast of prey, seems to have been evoked by Freud's first interpretation of his hatred towards his mother, who in fact dominated his life. The memory of his *father* evoked by this seems also to have acted as a screen, steering Freud away from this line of enquiry. Freud's interpretation was given in response to the Rat Man's associations which pictured Freud's own mother dead, with the Rat Man's Japanese swords stuck through her breast, and her genitals eaten into by Freud and his children like the rats of his phantasy. Such material, as it appeared regularly in the play of children later in the history of psychoanalysis, was to become the basis of Kleinian inferences about the primary role of hatred towards the mother and her breast, as shown in attacks in phantasy with all kinds of weapons. In Hopkins (2003, 2004), I describe how (I think) this original repressed aggression can be seen as the origin of that shown in outgroup conflict, as perpetuated by processes of group selection.