S & CB (2002), **14**, 143–158 0954–4194

NANCEY MURPHY

The Problem of Mental Causation: How Does Reason Get its Grip on the Brain?

1. Introduction[a]

Twenty years ago, when I first became involved in the theology and science dialogue, it was possible to ask whether there was really anything for scientists and theologians to talk *about*. It is important to remember that some of the most powerful influences in the development of modern theology, such as Immanuel Kant and Friedrich Schleiermacher, had argued that religion and science have nothing to do with one another.

Various intellectual strategies for insulating theology from science have made use of body–soul dualism. Put crudely, science can study the body but the soul is the province of theology. Such strategies, however, have become problematic in that neuroscientists are now studying all of the human faculties once attributed to the soul.

I would argue that Christians who have not already done so ought to join philosophers and neuroscientists in adopting a physicalist account of the person.[1] The problems with dualism, in my judgment, are insurmountable. First, it may well be conceptually impossible to give an account of mind–body interaction: how can something non-material interact causally with material entities? Second, while neuroscience can never prove that there is no mind or soul, it is increasingly clear that, to quote Laplace out of context, we have no need of that hypothesis. Finally, in addition to being *unnecessary* on biblical or theological grounds, dualism is theologically *undesirable* due to its penchant for distorting Christian priorities. Briefly, what I mean here is that the adoption of dualism gave Christians something to care about (their souls) in place of Jesus' primary concern, which was the Kingdom of God.

There are problems with physicalism, also. Most of the problems come down,

a This paper was first given as a lecture jointly organised by Christians in Science and St. Edmund's College at Trinity College, Cambridge University, on 8th November 2001, kindly supported by the Templeton Foundation.

1 There is controversy over the proper term to use for an account of the person that denies that we are composites of body and something else. 'Physicalism' is the term most often used by philosophers who hold this view.

in one way or another, to the issue of reductionism.[2] If humans are essentially bodies, can we still understand ourselves to have features once attributed to an immaterial mind or soul, such as rationality, morality and free will?

What I intend to do in this paper is to suggest the outlines of an approach to the problem of rationality. Here is the problem in brief: if humans are purely physical entities, how can it *fail* to be the case that their thoughts are determined by physical laws and, if so, what happens to our conception of rationality? This problem is discussed in the philosophical literature as the problem of mental causation. Philosopher of mind Jaegwon Kim expresses it as a dilemma: he argues that mental properties will turn out to be reducible to physical properties unless one countenances some sort of downward causation. But such downward efficacy of the mental would suggest an ontological status for the mental that verges on dualism.[3] My plan in this paper is to suggest a strategy for solving the problem of mental causation by providing an account of the downward efficacy of the mental that leaves an ontologically physicalist account of the human person intact.

2. Problems with nonreductive physicalism

The problem confronting the physicalist can be evoked using a simple diagram. See Figure 1.

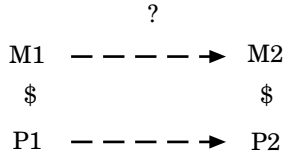

Figure 1

Here M_1 and M_2 represent mental states or properties; P_1 and P_2 represent physical states or properties.[4] The arrow from P_1 to P_2 represents a causal relation, and I use the dollar sign to represent the supervenience relation. The con-

2 'Reductionism has a variety of related meanings. Methodological reductionism is a strategy in science that seeks to understand an entity by studying its parts. Epistemological or theoretical reductionism is the thesis that sciences above physics in the hierarchy of the sciences can or should be related to lower-level sciences by means of definitions or bridge laws. The important issue here is causal reductionism, the thesis that the behavior of an entity is determined by the behavior of its parts.

3 See Jaegwon Kim, 'The Myth of Nonreductive Materialism' in Richard Warren and Tadeusz Szubka (eds.) *The Mind–Body Problem,* Oxford: Basil Blackwell (1994), pp. 242–60.

4 The relations among events, properties or states of events or entities, and descriptions all need to be worked out more carefully. I shall not attempt to do so here. I merely note that there may be a problem with Kim's use of these terms. For Kim, an event is the instantiation of a property at a time. But if there is a supervenient mental property and a physical property, both instantiated at t, how are we to know what is the relation between M1 and P1? Is it identity? If so, we seem to face

cept of *supervenience* is widely used in philosophy of mind to describe the relation between mental events or properties and brain events or properties. It was first employed to describe the relation between moral and descriptive attributes, and so we might call it a constitutive or 'in virtue of' relation. St. Francis has the moral property of goodness in virtue of his particular set of actions and character traits, and hence the property of being good supervenes on these non-moral properties. Similarly, it is hypothesized, one is experiencing mental event M_1 in virtue of undergoing a particular set of brain events, P_1. There is considerable controversy over an exact definition of 'supervenience' but for present purposes I need not pursue this.[5]

The diagram, then, expresses the assumption of causal closure at the physical level; that is, every physical event (in this case, the neurobiological event P2) has a sufficient physical cause. What, then, of the relation between the mental events, represented by the dashed arrow and the question mark?

The dilemma for the nonreductive physicalist comes down to this. Mental properties can be taken to have causal efficacy in so far as they supervene on physical properties and those subvenient physical properties are causally efficacious. But if the physical properties are causally efficacious, what causal work is left for the mental properties? We seem to be left with a new version of epiphenomenalism. Thus, I intend here to sketch out the basics of an argument for the compatibility of reasoned connections at the mental level with causal connections at the neurobiological level. To do so I shall turn, eventually, to the concept of downward causation.

First, let me make it clear that I am reframing Kim's question. Kim speaks in terms of mental and physical *properties* of events: if the physical property is causally sufficient, what is left for the mental property to do? I want to argue that this way of describing the problem misses the crucial issue. The crucial issue is whether the sequence from M_1 to M_2 is a *reasoned* sequence or merely a *causal* sequence. So, for example, I say to you '5 times 7.' You all think '35'. Did that happen because it's *true* that $5 \times 7 = 35$ or because a causal process in your brain made you think it?

---

all of the well-known problems with the various versions of the mind–brain identity thesis. If it is mere correlation (as Kim's definitions of supervenience would suggest) then we seem to have psychophysical parallelism or a new version of epiphenomenalism *unless* M1 is hypothesized to play a causal role in producing M2/P2. But this looks suspiciously like mind–brain interactionism. So I am inclined to emphasize that there are only two events here, neutrally referred to as e1 and e2, but susceptible of both mental and physical descriptions. To argue this, though, I would need to spell out, first, an account of supervenience different from Kim's, and second, deal with the problem of the relation between properties and descriptions.

5 Supervenience is understood by Kim and many others in terms of co-variation of properties: there can be no mental difference without a physical difference. Let me insert a demurral here: this account of supervenience does not seem to me to capture the original sense of the term – the *dependence* of the supervenient on the subvenient – and it *does* seem to ensure reducibility of supervenient properties.

Given that we presuppose the truth of 5 × 7 = 35, that is, that it is *rational* to think '35' when I say '5 times 7', we can again reframe the question: how can we reconcile an account in terms of reasons with a physicalist account of the mental without giving up on the causal closure of the physical? Colin McGinn asks: 'How, for example, does *modus ponens* get its grip on the causal transitions between mental states?'[6] I would rephrase his question as follows: 'How does *modus ponens* get its grip on the causal transitions between *brain* states?'

3. Some resources from philosophy and neuroscience

So I have identified the question of mental causation with the question of whether sequences of events (events simultaneously mental and neural) fall into patterns that are recognizably rational. And if so, how can we reconcile this with the supposed causal closure of the physical world? A hint about where I am going in this paper: notice that a calculator obeys the laws of physics *and* the laws of arithmetic. This is because it has been built in such a way that its causal processes model arithmetic transformations. The calculator has been *structured* in such a way that (any token instance of) a series of *triggering* causes – pressing the '5' key, the 'times' key, the '7' and the 'equals' – causes the machine to display a '35'.

Fred Dretske has introduced a distinction between triggering and structuring causes, which he illustrates by means of the following example: 'A terrorist plants a bomb in the general's car. The bomb sits there for days until the general gets in his car and turns the key. … The bomb is detonated (triggered by turning the key in the ignition) and the general is killed.' The terrorist's action was the structuring cause, the cause of its being the case that turning the key sets off the bomb.[7]

So for many purposes it is an oversimplification to represent a causal sequence simply as a series of events: E1 → E2 → E3. Instead we need to think of *two* series of events: those leading up to the triggering of the effect as well as those leading up to the condition under which T is able to cause E. Figure 2, adapted from Dretske's diagram, is intended to represent these intersecting strings of triggering and structuring causes:

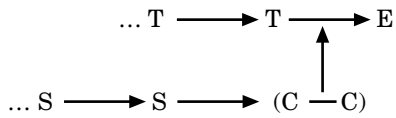

Figure 2

6 Colin McGinn, 'Consciousness and Content' in Ned Block, Owen Flanagan, and Güven Güzeldere (eds.) *The Nature of Consciousness: Philosophical Debates*, Cambridge: Cambridge University Press (1997), pp. 255–307; 305.

7 Fred Dretske, 'Mental Events as Structuring Causes of Behavior' in John Heil and Alfred Mele (eds.) *Mental Causation*, Oxford: Clarendon Press (1995), pp. 121–36; 122–3.

Here the Ts represent a series of triggering causes and the Ss represent a series of structuring causes leading to the ongoing condition C such that T is able to cause the effect E.

We shall have come a long way toward the goal of this paper if we can explain how the brain becomes structured so that, in the happy case, causal processes model rational relations. So the question is, are there significant enough analogies between the human brain and a calculator such that we can plausibly assume that the 'wetware' has been structured in such a way that its causal processes model or instantiate rational sequences? The disanalogy, of course, is that the calculator has been intentionally designed by a rational agent. Can we provide a plausible account, based on what we now know of neurobiology, as to how such rational structuring might occur without having to presuppose rational agency? I believe that the answer is yes. But here my account necessarily becomes somewhat speculative due to the tentativeness and incompleteness of neuroscientific explanations.

The physicalist assumption is that a mental event, such as thinking of the number 5 or thinking of Grandma, supervenes on a neural event. I chose the thought of Grandma as an example in order to allude to a controversy in neuroscience. The question was whether it was reasonable to assume that brains come equipped with individual neurons designated for recognizing patterns – that is, whether there is a 'grandmother neuron' devoted to recognition of this one particular elderly woman, and other cells for each pattern that the brain is able to distinguish.

It is now believed that most recognition tasks depend on the activation of large nets or assemblies of neurons rather than on the firing of individual neurons. Here is an account by neuroscientist Alwyn Scott:

> [L]et us consider what may appear to be a simple memory: that of your grandmother. Most of us are conscious of our grandmothers. But how? What series of neural links, of connections to connections, allow us to conjure up those dear old dames?
>
> Presumably an image of grandma may come to mind, and that would appear to involve the optic lobes. But other parts of the memory relating to voice would have originated, presumably, in the temporal lobes. These recollections are connected to others related to things she said and did, the way her house smelled on Thanksgiving Day, the colors of her kitchen, and so on. Because the memory of your grandmother is no doubt imbued with emotional overtones, those cells, whose locations are not known, would also need to be activated. And finally there is the not inconsequential linguistic task of matching the word 'grandmother' to one elderly or even long-deceased human female who happened to be the mother of one of your parents. It is difficult, if not impossible, to see how a single so-called 'grandmother' cell would manage to bind all of these components of

a complex memory together.[8]

Thus, even the simplest of mental events probably supervenes on the activation of a vast network of interconnected neurons. The concept of a 'cell assembly' was introduced by Donald Hebb, and its formation is described as follows: 'Any frequently repeated, particular stimulation will lead to the slow development of a "cell-assembly", a diffuse structure comprising cells... capable of acting briefly as a closed system….'[9]

It is in the *training* of such assemblies that we begin to see downward causation. It is better described as downward causation from the environment to the brain rather than mental causation, but insofar as intentionality or reference is an essential ingredient in rationality we have here the beginnings of an account of the rational *structuring* of the brain. Before pursuing this line of thought, however, we need to explore the concept of downward causation, a concept that Arthur Peacocke has made familiar in the theology and science literature.

4. Excursus: defining downward causation

There has been a developing literature on downward or top-down causation over the past 40 years. Philosophical theologian Austin Farrer was clearly groping for such a concept in his 1957 Gifford Lectures. Seeking a way to argue 'that higher-level patterns of action... may do some real work and thus not be reducible to the mass effect of lower-level constituents', he says that 'in cellular organization the molecular constituents are caught up and as it were bewitched by larger patters of action, and cells in their turn by the animal body'.[10] Farrer's metaphor of higher-level organizations bewitching the lower-level constituents is the sort of talk that deepens the mystery rather than clarifies it.

Psychologist Roger Sperry sometimes speaks of the properties of the higher-level entity or system *overpowering* the causal forces of the component entities.[11] However, elsewhere Sperry refers to Donald Campbell's account of downward causation. Here there is no talk of bewitching or overpowering lower-level causal processes, but instead a thoroughly non-mysterious account of a larger system of causal factors having a selective effect on lower-level entities and their causal effects. Campbell's example is the role of natural selection in producing the efficient jaw structures of worker termites and ants:

8 Alwyn Scott, *Stairway to the Mind: The Controversial New Science of Consciousness*, New York: Springer-Verlag (1995), p. 78.
9 Quoted by Scott, *op. cit*., p. 81.
10 Austin Farrer, *The Freedom of the Will* (the Gifford Lectures, 1957), New York: Charles Scribner's Sons (1958), p. 57.
11 Roger W. Sperry, *Science and Moral Priority: Merging Mind, Brain, and Human Values*, New York: Columbia University Press (1983), p. 117.

> Consider the anatomy of the jaws of a worker termite or ant. The hinge surfaces and the muscle attachments agree with Archimedes' laws of levers, that is, with macromechanics. They are optimally designed to apply maximum force at a useful distance from the hinge…. This is a kind of conformity to physics, but a different kind than is involved in the molecular, atomic, strong and weak coupling processes underlying the formation of the particular proteins of the muscle and shell of which the system is constructed. The laws of levers are one part of the complex selective system operating at the level of whole organisms. Selection at that level has optimised viability, and has thus optimised the form of parts of organisms, for the worker termite and ant and for their solitary ancestors. We need the laws of levers, *and organism-level selection*... to explain the particular distribution of proteins found in the jaw and *hence* the DNA templates guiding their production.[12]

Downward causation, then, is a matter of the laws of the higher-level selective system determining in part the distribution of lower-level events and substances. 'Description of an intermediate-level phenomenon,' he says, 'is not completed by describing its possibility and implementation in lower-level terms. Its presence, prevalence or distribution (all needed for a complete explanation of biological phenomena) will often require reference to laws at a higher level of organisation as well.'[13]

Campbell uses the term 'downward causation' reluctantly. If it is causation, he says, 'it is the back-handed variety of natural selection and cybernetics, causation by a selective system which edits the products of direct physical causation'.[14]

We can represent the bottom-up aspect of the causation as in Figure 3:

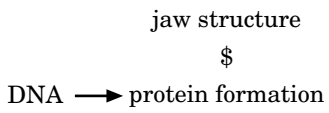

Figure 3

That is, the information encoded in the DNA contributes to the production of certain proteins upon which the structure of the termite jaw supervenes. This is micro-physical or bottom-up causation.

However, to represent the top-down aspect of causation, we need a more complex diagram, as in Figure 4, representing feedback from the environment,

12 Donald T. Campbell, '"Downward Causation" in Hierarchically Organised Biological Systems' in F.J. Ayala and T. Dobzhansky (eds.) *Studies in the Philosophy of Biology: Reduction and Related Problems*, Berkeley and Los Angeles: University of California Press (1974), pp. 179–186; 181.
13 *Ibid*., p. 180.
14 *Ibid*., pp. 180–1.

E. Here the dashed lines represent the top-down aspects, solid lines represent bottom-up causation.

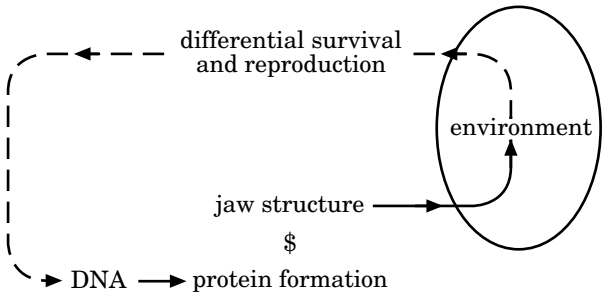

Figure 4

The most helpful recent account of top-down causation is Robert Van Gulick's.[15] Van Gulick makes his points about top-down causation in the context of an argument for the nonreducibility of higher-level sciences. The reductionist, he says, will claim that the causal roles associated with special-science classifications are entirely derivative from the causal roles of the underlying physical constituents of the objects or events picked out by the special sciences. Van Gulick replies that although the events and objects picked out by the special sciences are composites of physical constituents, the causal powers of such an object are not determined solely by the physical properties of its constituents and the laws of physics, but also by the *organization* of those constituents within the composite. And it is just such patterns of organization that are picked out by the predicates of the special sciences. Another way to make the same point is to say that physical outcomes are determined by the laws of physics together with initial and boundary conditions. Thus, Van Gulick concludes, 'we can say that the causal powers of a composite object or event are determined in part by its higher-order (special science) properties and not solely by the physical properties of its constituents and the laws of physics' (p. 251). The patterns of boundary conditions picked out by the special sciences have downward causal efficacy in that they can affect which causal powers of their constituents are activated or likely to be activated.

> A given physical constituent may have many causal powers, but only some subsets of them will be active in a given situation. The larger context (i.e. the pattern) of which it is a part may affect which of its causal powers get activated.... Thus the whole is not any simple function of its parts, since the whole at least partially determines what contributions are made by its parts. (p. 251)

15 Robert Van Gulick, 'Who's in Charge Here? And Who's Doing All the Work?' in Heil and Mele (eds.) *Mental Causation,* pp. 233–56.

Here we see a generalization of Campbell's insight that downward causation is not overpowering but selective activation of lower-level causal processes.

5. Downward causation of neural structure

Now, let us return to the question of how the brain becomes *structured* in such a way that its causal processes realize rational processes. Many theories of brain function rely on some form of 'neural Darwinism'.[16] That is, the answer to the question of how neural nets or cell assemblies form is by a process of random growth of dendrites and synaptic connections, followed by selective reinforcement of connections that turn out to be useful. The best theory seems to be that co-presentation of stimuli to two neurons or groups of neurons, resulting in simultaneous activation of their respective receptors, strengthens neuronal connections between those receptors, making it more and more likely that both cells or groups of cells will fire when one is stimulated. So useful connections (such as the connection between the 'grandmother' assembly and the 'cookies' assembly) are strengthened, while unused connections, (say, between 'grandmother' and 'frogs') weaken or die off. In this way, neural connections that model relations of various sorts in the world come to be selected.

If Campbell's account of environmental shaping of termite DNA is an instance of downward causation from the environment to the species' genome, then this, too, is downward causation. In this case it is from the environment to the individual brain during the individual's lifetime. A central claim of my paper, then, is that downward causation, in the sense of environmental selection of neural connections and tuning of synaptic weights, provides a plausible account of how the brain becomes structured to perform rational operations. In Van Gulick's terms, the larger system – which is the brain in the body interacting with its environment – selects which *causal pathways* will be activated.

So far we have an example of a weak form of rationality – presumably it *is* more rational to think of cookies than of frogs in association with thoughts of one's grandmother. Here the connections among things in the world come to be modeled by connections among cell assemblies in the brain. When this happens, free association is replaced by 'rational' trains of thought.

We can build from this beginning to consider more interesting forms of reasoning. If interaction with the physical world structures the brain in its image, so does interaction with the social world, with its structures and conventions. Consider the social environment of the primary school classroom, and the set of conventions we call arithmetic. How do the brains of children come to be structured so that neurobiological causal processes realize rational operations? Let us speculate about a simple form of learning such as rote learning of mul-

16 See, for instance, Gerald M. Edelman, *Bright Air, Brilliant Fire: On the Matter of the Mind*, New York: Harper Collins (1992).

tiplication tables. We can imagine that upon hearing the teacher say '5 × 7', neural assemblies are activated and, at first, activation spreads widely and randomly – activating a variety of other assemblies: for example, those subserving thoughts of, '57', 'Times Square', '30', '35', '75'. But feedback from the environment selectively reinforces one connection, while lack of reinforcement weakens all the others. We can picture this process by means of a diagram formally identical to the one I used to represent Campbell's account of downward causation. See Figure 5:

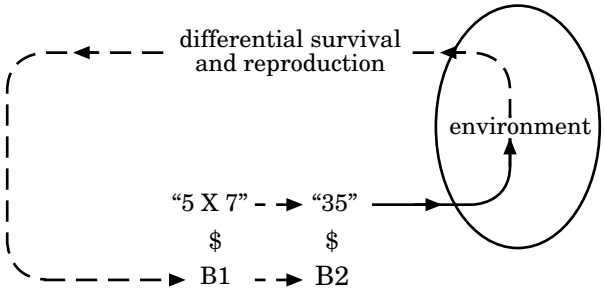

Figure 5

Here the thoughts of '5 × 7' and of '35' are pictured as supervening on two brain states (that is, activation of cell assemblies) labelled B1 and B2. Over time, feedback from the social environment results in a strong connection between B1 and B2 such that B1 regularly causes B2.

Let me emphasize that the foregoing is not intended to be a realistic cognitive-science account of the actual learning of arithmetic. In addition, it begs all of the questions pertaining to the foundations of mathematics. It is simply intended to show that downward causation in the form of environmental selection among neural connections provides a plausible explanation of how rational connections could become instantiated in or realized by causal pathways in the brain.

6. Downward mental causation

If my rephrasing of the problem of mental causation is satisfactory, perhaps this paper could simply end here. That is, I claimed that the issue is not the causal powers of the mental properties of events (as Kim says) but rather it is to explain how an account of a sequence of mental events ordered in terms of *reasons* can be reconciled with an account of those same events connected by neurobiological causes. But let us see whether we can elaborate the account I have given to provide not only for environmental shaping of the brain but for the reshaping of an individual's own neural pathways by means of one's own mental operations. This will give us clues to understand how rational norms in the brain itself (as opposed to merely the social environment) have downward

causal efficacy on the brain.

For this task we shall need to move to a higher level of abstraction – the information-processing level – in order to take account of the hierarchical structuring of cognitive processes. We shall also need a more detailed figure to represent the feedback processes involved. From neuropsychologist Donald MacKay I borrow Figure 6, a diagram for a simple feedback system:

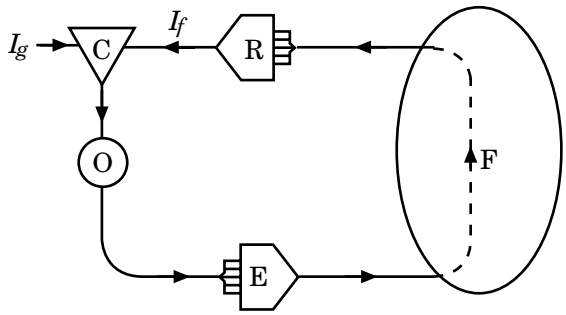

Figure 6

Here the action of the effector system, E, in the field of action, F, is monitored by the receptor system, R, which provides an indication, I_f, of the state of F. This indication is compared with the goal criterion, I_g, in the comparator, C, which informs the organizing system, O, of any mismatch. O selects from the repertoire of E action calculated to reduce the mismatch.[17]

This diagram can be used to represent any simple self-governing system, such as a thermostat. But it also allows us to represent more sophisticated cognitive operations. Consider again the learning of arithmetic. Instruction does not aim merely at rote learning but at teaching both skills and evaluation procedures that ultimately allow for internal correction of mental operations. In the case of rote memorizing of multiplication tables (as I described it above), the O in MacKay's diagram would represent the production of answers, at first, by random spreading of neural activation. But in the case of learning *how* to multiply, O represents a system that involves a skill or operation that produces answers. Feedback from the environment corrects wrong answers and at the same time trains and tunes the organizing system itself. Once this system has been trained it can take the place of the teacher, selectively reinforcing right answers and thus the neural connections that subserve them. So the brain becomes a self-modifying system, modifying its own neural structure in response to norms incorporated into its operations. The organizing system is trained by the environment and it in turn has downward causal efficacy in governing lower-level cognitive processes and thus the neural structures that subserve them.

17 Donald M. MacKay, *Behind the Eye* (the Gifford Lectures), ed. Valerie MacKay, Oxford: Basil Blackwell (1991), pp. 43–4.

But here we are still to some extent begging the question of mental causation because of the assumption that the norms according to which the organizing system operates are imposed by rational agents in the environment.

At this point I want to move from mathematics to logic since, at least for the rudiments, the relation between logic and experience is less of a philosophical minefield than the relation of mathematics to experience. Let me return to Colin McGinn's (rephrased) question of how *modus ponens* gets its grip on the causal transitions between brain states.

It is beyond my abilities to give a respectable cognitive-science account of how logic is learned, but MacKay's approach to cognition puts the emphasis first on the empirical rather than the a priori, and I shall follow his lead. Some sets of sentences or statements work in the world, in the field of action. Sets of sentences such as the following receive positive feedback from the environment. (1) If it is raining the streets will be wet. (2) It is raining. (3) The streets will be wet. Other sets, such as the following, result in negative feedback. (1) If it is raining the streets will be wet. (2) It is not raining. (3) The streets will not be wet.

These processes can be represented using Figure 6: the organizing system, O, operates on sentences in a natural language and makes a prediction that leads to action. A goal, such as keeping one's feet dry, is either met or frustrated. If the goal is frustrated, the comparator, C, makes a correction in O for next time.

Now, to know what 'modus ponens' means is to have acquired the ability to recognize a particular *pattern* among sentences. This ability allows for a more efficient correction of O. We now need a more complex figure to represent the process. In Figure 7 a supervisory system has been added. Rather than correcting O sentence by sentence, SS gives O general instructions to act only on sets of sentences that fit the pattern ((p —> q) & p) —> q (that is, If p implies q, and p, then q).

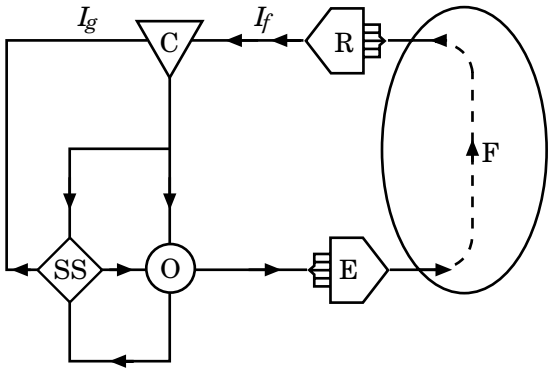

Figure 7

This diagram suggests that second-order cognitive operations are still influenced by feedback from F, only the feedback is indirect: for example, if SS instructs O to operate with the pattern ((p —> q) & -p) —> -q (that is, if p implies q, and not-p, then not-q) there will be massive failures to achieve goal states and as a result SS will be required to re-adjust O. Such a process is better represented in Figure 8:

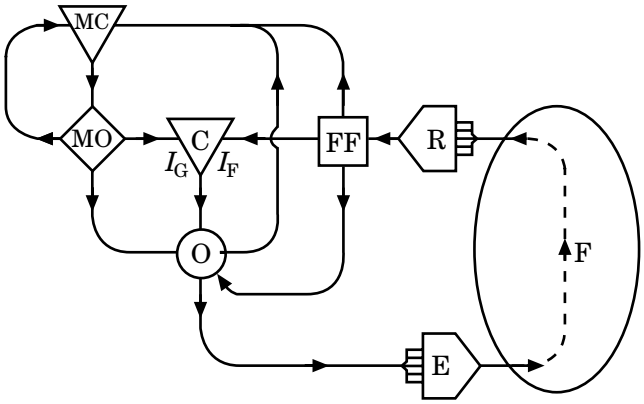

Figure 8

Here the supervisory system of Figure 7 is represented by two components, the meta-comparator, MC, and the meta-organizing system, MO. (FF represents a feedforward system with feature filters.)

In Figure 6, I_G, the goal state of the system is set by some criterion outside the system. For example, the goal of getting right answers in class is (initially, at least) set by the teacher. In the system represented in Figure 8 the goal state itself is set by higher-level processes within the system. MacKay says: 'we have drawn the meta-organizing system with a meta-evaluative procedure to take stock of how things are going. It adjusts and reselects the current goal in the light of its evaluation of the success of ongoing agency, and it also keeps up to date the organizing system to match the state of the world indicated by the feed forward from the receptor system' (p. 141). 'So there is a meta-organizing system, which organizes the activity of the organizing system in the way that the organizing system organizes the activity in the outside world' (p. 142).

The process of adding higher-level supervisory systems has no limit. We can use Figure 9 to represent the development of a system of logic in which *modus ponens* can be proved; SS now represents that system.

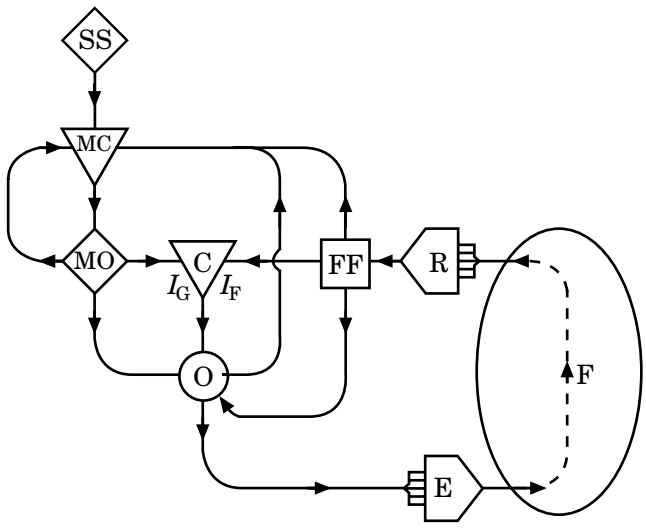

Figure 9

Notice that in the move from simple associative learning to higher levels of computation I have had to move from a rather straightforward model of neural processes to the level of cognitive-science description. However, I believe that research within the connectionist paradigm offers hope for describing in neurobiological terms the formation and training of these various systems that, for the present, must be referred to in abstract terms simply as 'organizing systems', or 'supervisory systems'.

7. Conclusion

My strategy in this paper was to bring together accounts of physical causation more complex than the ones that typically appear in philosophy of mind with a little knowledge from neuroscience and cognitive science about what is happening in the brain when we perform mental tasks.

What I hope to have shown is the following. First, causation in the physical world is not best understood in terms of linear sequences of events. We need analyses such as Dretske's that emphasize the role of physical structures in co-determining the effects of events. We especially need the concept of downward causation, which includes recognition of the existence of higher-level systems with their capacities to select among the lower-level causal processes upon which the system depends.

Second, I hope to have given enough of a sketch of neural processes to make plausible the claim that downward causation is pervasive here, from the structuring of neural nets by interaction with the environment to the construction of higher-order internal supervisory systems that, once in place, govern lower-

order cognitive operations.

In sum, through interaction with the environment and through development of higher-order feedback loops, neural networks become structured in such a way that their causal connections model (or instantiate or realize) rational connections among mental events.

Does this solve Kim's dilemma? His claim is that one cannot account for the causal relevance of mental properties of events without violating the causal closure of the physical. Let us return to the simplest of my examples, the thoughts of Grandma and of cookies supervening on brain events B1 and B2 respectively. In this example, being the thought of Grandma is the mental property; being the activation of a particular cell assembly is the physical property. What I hope to have shown is that Kim's challenge cannot be assessed if we consider only a token instance of this event and its relation to B2. We need to consider the history that has led to the strong neural connection between B1 and B2 in order to see the causal relevance of the mental. That is, it is because B1 realizes the thought of Grandma and B2 the thought of cookies that this strong causal connection has come to exist. Yet the laws of neurobiology were never violated in the process by which the environment selected this connection from among billions of other possibilities.

I acknowledge that this paper has barely scratched the surface of the task I set for myself, which was to reconcile an account of supervenient mental events obeying principles of reason with an account of brain events connected by neurobiological causation. A more satisfactory treatment calls for more realistic accounts of learning and other cognitive processes, and, of course, for connections between cognitive science and neuroscience that are not yet in place. Finally, to avoid the charge that I have simply substituted environmental determinism for neurobiological determinism I will need to deal with both nonalgorithmic reasoning and free will. I hope to do much of this in the future. For now, I look forward to your comments on this small piece of the puzzle.

Nancey Murphy is Professor of Christian Philosophy at the Fuller Theological Seminary, Pasadena, California, USA.