

Warren S. Brown, Fuller Integration Lectures

Lecture 3: “Did My Neurons Make Me Do It?”

February 18, 2005

I. Retrospect and Prospect

Let us begin by summarizing what we have covered in the first 2 lectures, and then set the task for today’s lecture. We have covered the following 4 issues:

1. *Is dualism a necessary part of a Christian view of humankind?* In Lecture 1, I suggested (without going deep enough to prove the point) that body/soul dualism has extra-Christian roots and is not a necessary presupposition for understanding persons in a way that is consistent with Biblical revelation. This was done to give us theological “elbow room.”
2. *What are the practical consequences of dualism for Christian praxis and for psychological theories and interventions?* In Lecture 1 I also critiqued both practical Christian theology and psychological formulations of the person with respect to the impact of a dualist understanding of persons. Absence of a wholistic and physicalist view of persons makes spirituality inward and individualistic, distances persons from their behavior, and weakens community.
3. *Is dualism tenable within the context of modern neuroscience and neuropsychology?* Out of data from recent research in neuroscience, neurology, and neuropsychology, I attempted in Lecture 2 to show why it is increasingly difficult to maintain the dualist idea that the most spiritual and soulish aspects of personhood are outside of the purview of the functioning of our brains.
4. *What might a monist (or physicalist) Christian theological anthropology be like?* In Lecture 2, I also ventured my suggestions regarding an evangelical theological

anthropology. I argued for a physicalist view of human nature that defined our humanness in relational terms – relationships to one another, to ourselves, and to God.

In this lecture I will be considering the following question: *How is it possible for physicalism to avoid reductionism and biological determinism? Is it possible to give a reasonable account of free will within a physicalist understanding of human nature?*

I will try to defend the position that physicalism can be understood in a non-reductive way – that is, in a way that does not presume that all humanness can be reduced to “nothing but” neurophysiology or the laws of physics, or that human behavior is entirely determined by physical laws. I will argue that we are causes of our own behavior via our mental processes, and that this “mental causation” (as it is often called) allows for free will, moral responsibility, and genuine personal relatedness.¹

II. The Problem of Reductionism and Determinism

Nancey [Murphy] expresses, in the mode of philosophical discourse, what is at stake in the problem of mental causation and moral agency as follows: “If mental events are intrinsically related to ... neural events, how can it *not* be the case that the contents of mental events are ultimately governed by the laws of neurobiology? If neurobiological determinism is true then it would appear that there is no freedom of the will, that moral responsibility is in jeopardy, and ... that our talk about the role of reasons in any intellectual discipline is misguided.”²

¹ This entire lecture is out of work done colloaboratively with Nancey Murphy for the following book currently in preparation: Nancey Murphy and Warren S. Brown. *Did My Neurons Make Me Do It? Philosophical and Neurobiological Perspectives on Moral Responsibility and Free Will.*

² Nancey Murphy, Supervenience and the downward efficacy of the mental: A nonreductive physicalist account of human action. In R.J. Russell, N. Murphy, T.C. Meyering, and M.A. Arbid (eds.) *Neuroscience and ther Person: Scientific Perspectives on Divine Action.* (Vatican City State: Vatican Observatory, 1999), 147.

Philosopher Fred Dretske expresses the task of giving an account of mental causation in this way: “The project is to see how reasons – our beliefs, desires, purposes, and plans – operate in a world of causes, and to exhibit the role of reasons in the *causal* explanation of human behavior. In a broader sense, the project is to understand the relationship between the psychological and the biological...”³

The questions on which we will focus in this lecture are the following:

- If we are physically embodied beings, is it inconsistent or incoherent to say that we consciously and willfully cause our own behavior?
- In what ways are biological systems self-causing agents?
- How does a complex, physical organism become a morally responsible human being?

III. Moral Agency without “The Soul”

Establishment of a robust psychological and theological understanding of human moral agency and free will within a physicalist understanding of the person is far from a simple, one-step (or one-argument) process. From the adoption of the hypothesis of monism or physicalism there is a line of argument that needs to be successfully made to be able to maintain a richly human and Christian view of the person. For starters, we must be able to imagine plausible solutions within physicalism that can lay to rest concerns about reductionism and determinism.

A very simple outline of this line of argument would be the following:

First, mental causation needs to be demonstrated. That is, we to establish how it can be the case that our mental life is the cause of our behavior. It needs to be shown how top-down causal influences (from mental processing to physical activity) can exist,

³ Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, Mass.: Bradford Books, 1988), x.

such that human mental activity and consequent behavior cannot be entirely explained by the more micro-level processes of our physiology.

Second, it needs to be shown how moral agency comes to be an aspect of mental causation. Together, these first two steps (mental causation plus moral agency) amount to a defense of a Christian understanding of free will within physicalism.

Finally, it needs to be shown that physically embodied human persons can be genuinely relational – that is, that out of a physically embodied person can come genuine manifestations of the relational ‘fruits of the spirit’ – love, joy, peace, patience, kindness, gentleness, and self-control – as well as “works of the flesh” – strife, jealousy, anger, quarrels, dissension, factions, and envy.⁴

Therefore, in this paper I will be sketching the bare bones of a physicalist, but non-reductive, understanding of human mental function that I believe can account for robust forms of human agency. I will propose a *biologically plausible* description of how mental properties with causal roles might emerge from the physical activity of hyper-complex human brains, and how moral agency might emerge from causally efficacious mental function.

IV. Views that Limit the Imagination

I talked in the first lecture about the important role of Rene Descartes in establishing dualism within Western thought patterns. Descartes is most responsible for the huge impact of body / soul dualism on modern secular and Christian thinking. I also described Descartes as primarily a physicalist in that he did NOT believe that the body was inhabited by many souls, as was commonly believed in Descartes’ time. Rather, Descartes believed that all basic bodily functions were aspects of a physical “machine”, and that the functioning of animals did not transcend these mechanisms. The problem

⁴ Galatians 5:19-23

for Descartes was figuring out how such a biological mechanism could result in human consciousness, will, and rationality. So, Descartes solved this problem by retaining *one* soul. Thus, humans were considered to be unique in having a soul (only one) that was immaterial and interacted with the physical body through the pineal gland.

The question I asked in that lecture was what might Descartes have concluded had he had before him the current body of neuroscience literature – for example, the fMRI studies that I described in my lecture yesterday? Descartes' imagination regarding the possibilities for physicalism was limited by a lack of sophisticated neuroscience. He could not have concluded otherwise.

My reason for repeating this account of Descartes is to point out that there are other limitations to our imaginations regarding the possibilities of a monist, wholist, or physicalist view of human nature. Two thought-paradigms have been superimposed on human behavior and neurobiology that have come to strongly influence our understanding of human nature and limit our imaginations.

One thought paradigm is the reductionist view of causes that has been brought to biology from physics. In this view, all the causal work in the physical world is done at the level of atomic and subatomic matter. This idea has been an important source of reductionism in the views of many neuroscientists, and has fostered what Donald MacKay has called “nothing-buttery” – that is, our mental life is “nothing but” the operation of micro-level physiology and physics. From physics we have also inherited an aversion to the idea of self-cause. According to the world view of physics, an object within the physical world cannot be the cause of its own activity.

Since all the forces available for consideration are subatomic, and since the physical world is a causally closed system (as the story goes), “moral behavior” is logically incoherent because it can only be the outcome of physical forces at the subatomic level. Thus, for the person with an eye to moral agency, personal

responsibility, or the spiritual potential of persons, the obvious conclusion can only be Descartes' solution – add a nonmaterial part. Without a richer imagination for the potentialities of biological systems, one is left with the limited choice between a *meaningless* human nature determined by subatomic forces, or an enriched and *meaningful* human nature that is endowed by a numinous, non-material soul.

Another limitation on the imagination comes from the philosophy of mind. The limitation here comes from the abstract and linear conceptual paradigms often used in philosophical discussions of mind. This simple and linear mode of thinking takes the form of diagrams indicating *physical states leading to physical states*, with *mental states* attached to physical states. The question is whether the mental states have *any* causal role in themselves. This sort of thinking caused philosopher Jaegwon Kim to reject the possibility of a *nonreductive* physicalism.⁵ Chains of physical causes are, within this limited logic, considered sufficient – adding the mental as a cause merely results in multiple determination of actions.

In my opinion, such simple linear models have almost nothing to do with physiology and, therefore, the rejection of a view of mind based on this form of argument is misconstrued. The very formulation of the problem leads inescapably to either dualism or reductive materialism. When the imagination is captured by these linear, abstract, and simplistic (with respect to complexity of human brains) diagrams and syllogisms it is unlikely that one will come up with solutions that are relevant to the complexities of human biology and behavior.

⁵ See Jaegwon Kim, "The Myth of Nonreductive Materialism," in *The Mind-Body Problem*, Richard Warren and Tadeusz Szubka, eds., (Oxford: Basil Blackwell, 1994), 242-60.

V. Emergence and Top-Down Causation

There are two concepts that need to be understood and defended in order for our imaginations to be enriched regarding the possibilities for causally efficacious human mental processes: emergence and top-down causation (or top-down influence).

“Emergence” denotes the fact that *more complex entities* can have properties that do not exist within the *elements* that make up the complex entity. To expand on this important concept and to give it more detail, the following points are critical:

- New functional capacities arise from the interactions between the parts of a complex system.
- Interactions create larger dynamic patterns that bind smaller units into webs of interactive influences.
- These larger dynamic patterns operate by different rules and causal processes than the parts (i.e., they have emergent properties).
- What is more, the relational constraints of the larger pattern on the individual parts significantly increases the range of possibilities for the system as a whole.

Emergent properties have top-down influence on the activity of the constituent parts – our second important concept. The idea here is that micro-level processes (e.g., the physics and chemistry of neurons) become caught up in, and influenced by, the larger dynamic patterns that constitute mental events. Thus, the emergent phenomena of mental processes (including thinking, deciding, consciousness, memory, language, representation, belief, etc.) create top-down influences on the lower-level neurophysiological phenomena whose activity constitute the mental processes themselves.

The mechanism and exact form of this top-down influence is still being debated. It is perhaps most easily integrated with concepts from physics (e.g., the theories of a

casually closed system, conservation of energy, and determinism at the subatomic level) if one understands top-down influences to be whole-part constraints. The activity of the system as a whole constrains the possibilities at the micro-level without interference in the lawfulness of the local physical operations. Dynamic systems theory can give a good account of the emergence of causal properties in complex interactive systems based solely upon whole-part constraints on the behavior of the parts of the system. There is not time in this lecture to complete this point. I refer you to our book.⁶

The point here is that we need to be able to consider theories of emergence and top-down influence in order to make sense of the physicality of our humanness. In order to further enrich our imaginations with some data, let me relate the outcome of three experiments that I believe are illustrative of the reasonableness of these ideas.

First is an example of emergence, illustrating the importance of brain interconnectivity for the emergence of important human cognitive capacities. This example comes from my own research on agenesis of the corpus callosum (ACC).⁷ Comparison of the midline MRI views of a normal brain with that of an individual with ACC reveals the very obvious absence of a major brain pathway – the corpus callosum. The corpus callosum is composed of over 200 million nerve fibers that interconnect the 2 cerebral hemispheres. In ACC, this structure has simply failed to develop, significantly diminishing the interconnectivity of the right and left cerebral hemispheres.

Normal children between about 5 and 10 years of age develop the ability to understand non-literal and metaphoric language. The critical event in the emergence of this capacity seems to be completion of the development of these fibers that

⁶ Chapter 3, Section 3 of: Nancey Murphy and Warren S. Brown. *Did My Neurons Make Me Do It? Philosophical and Neurobiological Perspectives on Moral Responsibility and Free Will.*

⁷ Brown W.S. and Paul L.K., (2000) Psychosocial deficits in agenesis of the corpus callosum with normal intelligence. *Cognitive Neuropsychiatry*. 5, 135-157.

interconnect the two cerebral hemispheres. I know this to be the case because the individuals with ACC that we have studied have considerable deficits in understanding non-literal and metaphoric language (despite otherwise normal IQs).⁸ Therefore, the neural connections and interactions between the hemispheres via the corpus callosum must be necessary for the emergence of this very complex, high-level cognitive process – a process that does not emerge when the 200 million neurons of the corpus callosum are missing and the hemispheres must work independently.

The second illustration is a particularly clear and oft-replicated example of top-down causation from the “mental” to the “physical” in the form of the placebo effect. The placebo effect is created when higher cognitive processes of understanding and belief regarding the effectiveness of a sugar pill (for example) create a top-down effect (cognitive process-to-cellular systems) that has been repeatedly shown to relieve pain and/or enhance the activity of immune cells.

A recent neuroscience study reported functional brain imaging (or fMRI) demonstrating that a placebo, believed by the recipient to relieve pain, activates the very same brain structures that are activated by an injection of opiate drugs. In essence, belief (embodied in larger dynamic patterns of brain activity) controls pain by altering activity in the brainstem and anterior cingulate gyrus. The article concludes that this research “graphically illustrates the principle that higher brain functions help control how humans perceive pain.”⁹

Finally, another example of the impact of mental activity on the activity of specific brain regions comes from the work of Baxter and colleagues at UCLA. These

⁸ Paul, L.K., Van Lancker, D., Schieffer, B. and Brown, W.S. (2003) Communicative deficits in individuals with agenesis of the corpus callosum: Nonliteral language and affective prosody. Brain and Language. 85, 313-324,

⁹ P. Petrovic, E. Kalso, K.M. Petersson, & M. Ingvar. Placebo and opioid analgesia – Imaging a shared neuronal network. *Science*, 295, 1737-1740, 2002.

investigators were able to show that a cognitive-behavioral therapeutic process that involved having the patient consciously withhold obsessive-compulsive behavior – a *top-down process* – had both the same behavioral outcome (i.e., reduced obsessive compulsive behavior), and the same eventual effect on distributions of neural activity as the recommended drug (clomipramine) – a *bottom-up process*. That is, the cognitive-behavioral therapy and clomipramine both resulted in the same changes in activity of the caudate nucleus of the brain (as evident in PET scans).¹⁰

VI. Is the “Mental” in the Brain?

If the mind is embodied, it seems obvious that the mind is what the brain does. However, consider the following thought experiment described by Donald MacKay.¹¹ You imagine that you are sitting at a table with your own functioning brain sitting in front of you (still connected to your body and nervous system). In this thought experiment, you electrically stimulate your own brain in the primary visual area, and, of course, you will see flashes of light. MacKay believes that several issues can be clarified from this simple thought experiment:

1. *Where are the physical events that are creating your experience of lights flashing?*
They are in the part of the brain being stimulated. Thus, brain activity is necessary for your subjective mental experience of seeing lights flashing.
2. *Where are the lights as far as you are concerned?* They are out in your immediate action-space. They are part of your map of the external world – the field with which you are currently interacting. Mental events, even if dependent on internal physiological events, are about the field-of-action (or about you-in-the-field-of-action).

¹⁰ Baxter, L.R., Schwartz, J.M., et al. Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. Archives of General Psychiatry. 1992. 49, 681-689.

¹¹ Donald M. MacKay, *Behind the Eye*. (Oxford, UK. Basil Blackwell Ltd., 1991), 5-8.

Subjective mental states are about our action-space, even when we are being “introspective” and constructing an abstract self-referential action-space.

3. *But, Where are YOU?* You are sitting in front of your brain. You are not on the table in front of you. Your thinking and experiencing are attached to you sitting in the chair – not to the subpart of you sitting on the table. Therefore, all descriptions of human mental experience are about the whole embodied person, not about the brain as a separate organ. In this respect, the mental has to be something resident in the whole person.

Thus, MacKay’s thought experiment suggests that “mental states” are not internalized in the sense of existing in a form that is uncoupled from contexts of interaction with the real world. In its root form, a “mental state” is the cognitive resources being brought by the whole person to an ongoing interaction with the world. Mental activity is the entire *embodied* complex organism *embedded* in a process of interacting with the current field of action (real or imagined). It is not a description of the activity of a separate, inner planner and initiator of thoughts or behaviors. This is what “non-reductive” means – mental states are about the whole person contextualized within an action-space. The brain is necessary for mind, but not sufficient.

MacKay makes another important point concerning our use of any mentalist language with respect to human behavior and human brains. According to MacKay, we should attempt to exercise “semantic hygiene,” that is, we should only use mentalist language to describe what the *whole* person (or animal) is doing or experiencing, not to refer to the activity of some *part* of the person (such as the brain or a specific area of the brain).¹²

¹² MacKay, op. cit., 8-10.

VII. The Mental “Out and About”

Despite the fact that what we mean by “mental” is about the whole physical person interacting with the physical and social world, we must also recognize that a major contribution to the emergence of mind is made by phenomena that are *not* within the brain or body. Andy Clark argues for the importance of “external scaffolding” in the emergence of higher mental processing and human intelligence.¹³ “Scaffolding” refers to all of the ways that an organism relies on external supports for augmenting mental processing. Clark writes, “We use intelligence to structure our environment so that we can succeed with less intelligence. Our brains make the world smart so that we can be dumb in peace! ... It is the human brain *plus* these chunks of external scaffolding that finally constitutes the smart, rational inference engine we call mind.”¹⁴

The emergence of nonreducible properties of mind and the causal role of the mental becomes much less mysterious if we take into account the continued interactions between the person and the physical and social environments – interactions that involve creation and use of external structures to scaffold cognition. Our interactions with the physical and social world occur within an environment that is pre-prepared for supporting intelligent activity and that enriches the range of possibilities for mental causation. Human culture involves a vast array of artifacts that scaffold cognitive processing, the most remarkable of which is language. Scaffolding allows us successfully to solve problems that would exceed the capacity for immediate, on-line brain processing. It is also obvious, when we think about it, that social and cultural scaffolding provides us with concepts of morality and provide the basis for richer possibilities for human relatedness.

¹³ Andy Clark, *Being There: Putting Brain, Body, and World Together Again* (Cambridge, Mass.: Bradford Books, 1997), 179-192.

¹⁴ *Ibid*, p.180.

VIII. Mind as Action Loops

In order to understand the nature of mental causation, we first need to enrich our view of the general nature of the behavior of all organisms. Many of the causal diagrams used by philosophers of mind, as well as the stimulus-response models in psychology, give the impression that organisms are fundamentally passive, and behavior must be initiated (or triggered) either by external stimuli, or by an inner homunculus. I do not think that “triggering” represents a correct and informative way to view animal or human behavior.

It is more accurate biologically to view all organisms as characterized by three important properties:¹⁵

1. *All organisms are continuously active.* Behavior is emitted from inside the organism not triggered by external stimuli. Thus, behavior is always (or nearly always) voluntary, rather than elicited in a passive organism by a particular stimulus.

Martin Hiesenberg is a neuroscientist who works on the behavior of fruit flies (*drosophila*). He has studied flying maneuvers attempted by flies that are tethered to a strain gauge. Hiesenberg has demonstrated in his experiments that the flying behavior of fruit flies is a matter of *trying out* potential directions of flight in order to satisfy current internal needs. Heisenberg has demonstrated in his experiments that *no* external stimuli can be detected that *cause or trigger* attempted changes in direction.¹⁶ In this sense, the behavior is voluntary – that is, emitted, not triggered. All organisms are, by nature, continuously active, and, thus, all behavior is voluntary.

¹⁵ These characteristics are taken from a book in progress by N. Murphy and W. Brown, *Did My Neurons Make Me Do It? Philosophical and Neurobiological Perspectives on Moral Responsibility*.

¹⁶ Martin Heisenberg, “Voluntariness (Willkürfähigkeit) and the General Organization of Behavior”, in R.J.Greenspan and C.P.Kyriacou, *Flexibility and Constraint in Behavioral Systems* (John Wiley & Sons, Ltd., 1994).

2. *All behavior of organisms is goal-directed.* Behavior is tried out in order to accomplish a goal (whether these goals are represented simply or complexly within particular organisms).

Even protozoa have goals to find nutrients and avoid toxic substances. Like the fruit fly, they emit swimming behavior in order to try out directions of movement to determine if they will result in more nutrients or less toxicity. In more complex organisms capable of associative learning, there exist both basic biological drives (i.e., goals based on biological needs) and derived goals that are acquired through learning (particularly social learning). Thus, “goal-directed” can be relative to immediate biological needs, or to derived desires that are not immediately relevant to such needs.

3. *All organisms have the ability to evaluate the outcome of behavior and modify their ongoing behavior in relationship to evaluations.*

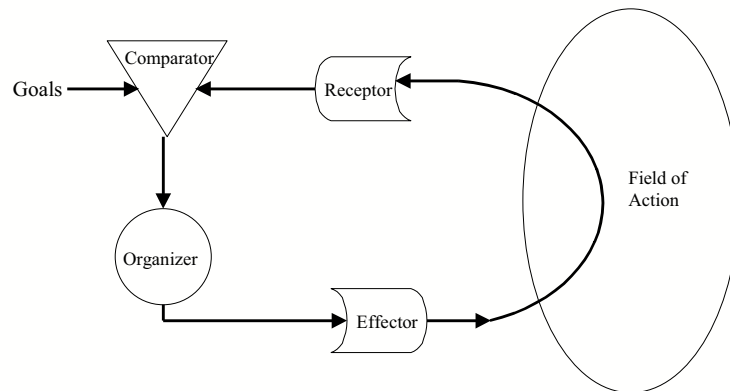
As we can already appreciate in the behavior of protozoa and drosophila (as well as laboratory rats, monkeys, and human beings), the activity of all organisms is a constantly recurring loop of emitting behaviors, evaluating the behaviors based on a comparison of sensory feedback and internal criteria for desired outcomes, and adjusting ongoing behavior based on the evaluation of feedback.

Donald MacKay in his book *Behind the Eye* (the content of his 1986 Gifford Lectures) presented an information-engineering diagram of the minimal functional architecture of an agent that captures these 3 aspects of the behavior of all organisms.¹⁷ In the flow diagram below, Effectors allow us to interact with the Field-of-Action. The outcome of our behavior is detected by our sensory Receptors. The outcome is compared to the Goals of the actions in the Comparator, where any mismatch would be fed to the Organizing System that would adjust the activity of the Effectors ... and so on. Thus, this diagram expresses the fact that behavior always involves a process of

¹⁷ Figures adapted from MacKay, *Behind the Eye*. p. 43.

continuous loops of action-feedback-evaluation-action (which I will call “action loops”). Behavior has no beginning or end, and is not typically triggered by external stimuli. Rather, feedback from the field of action is used to constantly monitor the success of the current goal-directed behavior of the organism with respect to desired outcomes, and necessary adjustments in behavior are made.

MacKay’s conceptualization of the minimum functional architecture of an agent



Consider the following example of high-level mental causation evidence only in the process of acting. The following is an account of an experiment regarding procedural knowledge done by Berry and Broadbent as it appears in John Anderson’s textbook entitled, Cognitive Psychology and It’s Implications:

Berry and Broadbent (1984) ... asked subjects to try to control the output of a hypothetical sugar factory (which was simulated by a computer program) by manipulating the size of the workforce. Subjects would see the month’s output ... and then choose the next month’s workforce. [The rule relating workforce and

output was a complex mathematical formula.] ... Oxford undergraduates were given sixty trials at trying to control the output of the sugar factory. Over the sixty trials, they got quite good at controlling the output of the sugar factory. However, they were unable to state what the rule was and claimed they made their responses on the basis of “some sort of intuition” or because it “felt right.” Thus, subjects were able to acquire implicit knowledge of how to operate such a factory without corresponding explicit knowledge.¹⁸

This experiment helps us focus on several aspects of action-loops, as well as the embodied and embedded nature of mental causation:

1. The behavior of these university students in interacting with the sugar factory program is clearly a form of *very* high-level mental activity, even though the students had no *conscious* awareness of exactly what they had learned and why they were eventually able to perform the task so well.
2. Knowledge regarding how to solve this problem only appeared in the interaction between the person (body and brain) and the computer program. While experience with the program obviously modified their brains in some manner, this brain modification was inert and only became “mental” as it was realized in interacting with the computer program. In this case, there was no access to the knowledge outside of engagement of a particular task-relevant action-loop.
3. This very high-level form of mental processing was embodied in unconscious criteria for evaluation of ongoing behavior.

This view of the behavior of organisms puts a different spin on the idea of a “cause” in mental causation – that is, “causation” is not triggering of action in an otherwise inert organism. It is not necessary to find the cause that set a particular

¹⁸ John R. Anderson, *Cognitive Psychology and Its Implications* (New York: Worth Publishers, 2000), 236-237.

behavior in motion. Rather, “causes” of behavior are the processes by which a continuously active organism evaluates and modulates its action. The causal role of the mental is evident in *evaluative modulations* of ongoing behavior, not in the initiation of behavior.

IX. The Problem and an Illusion of a Problem

Having established, I hope, that mind is an active process, it allows me to assert that the *problem* of mental causation, in the form it is often discussed, is an illusion. It is a residual phantom of a Cartesian view where the mind (or soul) is presumed to be interior and effectively disembodied. In a dualist view, mind is thought to be an inner autonomous agent or homunculus – as represented in the metaphor of a “Cartesian theatre.” Consequently, willful behavior cannot occur until moved by the inner mind. Otherwise, the person is largely passive.

I suggest the alternative possibility that “mind” is something that occurs in *acting*. At their core, mental processes are neither prior to, nor apart from, doing. To repeat a major point, mind is embodied (i.e., involving both the body and brain) and embedded (i.e., contextualized in action). As Andy Clark expresses, “minds make motions.” Mind is always “on the hoof” – contextualized in action.¹⁹ Thus, mind is not the brain, but neither is it a non-material emergent in some numinous and dualist sense. Rather, mind is a description of the brain and body operating as one in solving real problems in the field of action. In this sense, “mind” should always be a verb...we “mind” we do not have “a mind.”

So, what I meant when I said that the problem of mental causation is illusory is that mind is, in this definition, causes-in-process. Perhaps it would be more accurate to propose that mental causation poses a different sort of problem than many previous

¹⁹ Andy Clark, *Being There*, p. 1.

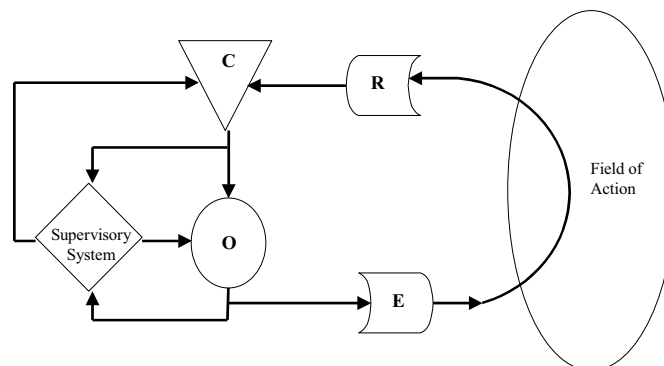
discussions have imagined, but it is the sort of problem that is comprehensible within the sort of nonreductive psychobiology that I have suggested.

X. Hierarchy and Complexity

I have been proposing a view of mind as resident in action loops, which, thus far, has been conceived in rather simple terms. However, a complete account of the causal role of the mental needs also to involve an understanding of more complex forms of mental processing. Yet even these more complex forms can be understood as elaborations of action-loops. Thus, we need to consider the overlays of more complex forms of modulation of action.

Here the diagrams of Donald MacKay are again useful. MacKay expands his functional architecture of an agent to suggest overlays of more complex and abstract forms of action modulation. He suggests that action loops are modulated by supervisory systems that function to set goals. This next diagram from MacKay incorporates the concept of a Supervisory System.²⁰

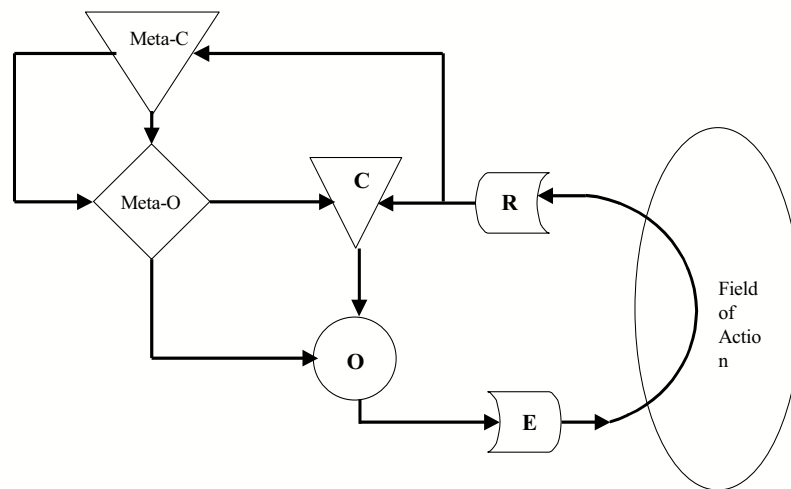
Functional architecture of an agent that sets its own goals



²⁰ Figures adapted from MacKay, *Behind the Eye*.p. 51.

MackKay makes it clear that this “supervisor” is not some sort of homunculus, nor even a centralized place in the brain where, in Daniel Dennett’s terms, “it all comes together.” Rather, a supervisory system is a larger action loop within which the original loop is nested. Therefore, this diagram when elaborated to illustrate the nature of a supervisory system looks like this:²¹

Functional architecture of a supervisory system: Nested hierarchy of organization

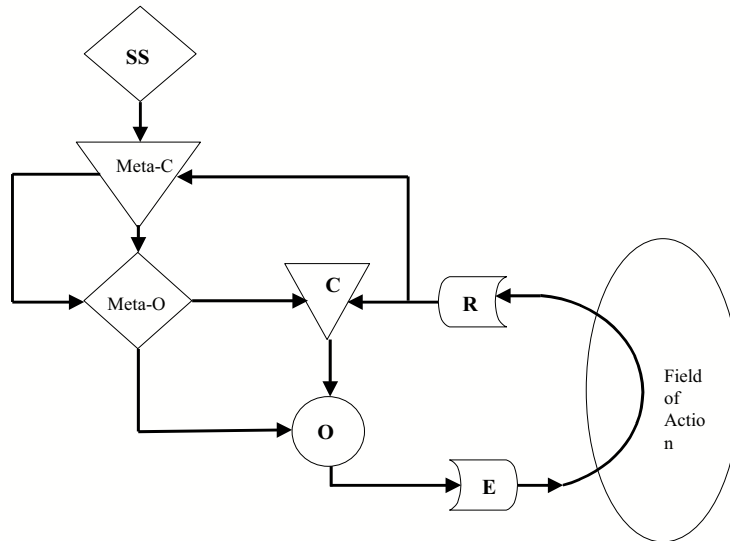


In this version of MacKay’s diagrams we see that the functional architecture of a supervisory system is a meta-comparator and meta-organizer – that is, the same sort of architecture that comprises the original action loop, involving action, feedback, evaluation, and modified action. MacKay points out that it is reasonable to consider

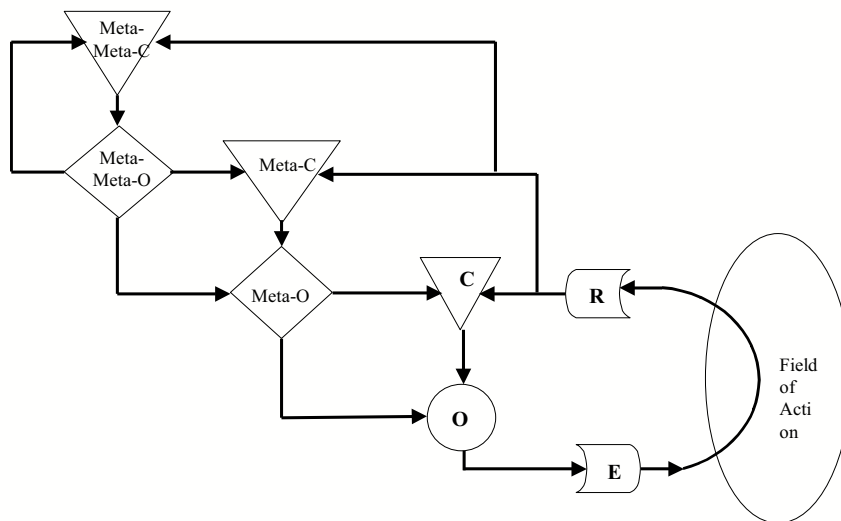
²¹ MacKay, op. cit., 141.

increasingly more complex levels of nesting of such modulatory feedback loops, as in the next diagram.²²

**Continued elaboration of a nested hierarchy:
A supervisor of the supervisory system**



**Even further elaboration of a nested hierarchy:
Meta-meta comparator and organizer**



²² Modification of MacKay, op. cit. p. 141.

Only the limits of imagination constrain the possibilities for nesting, particularly given the incredibly complex network that is the human brain. MacKay does not speculate as to what might be the various forms of comparisons and evaluations in the higher level loops, but one can imagine them to involve more complex memories, information distributed over longer epochs of time, and more abstract forms of representation eventually involving symbolic systems.

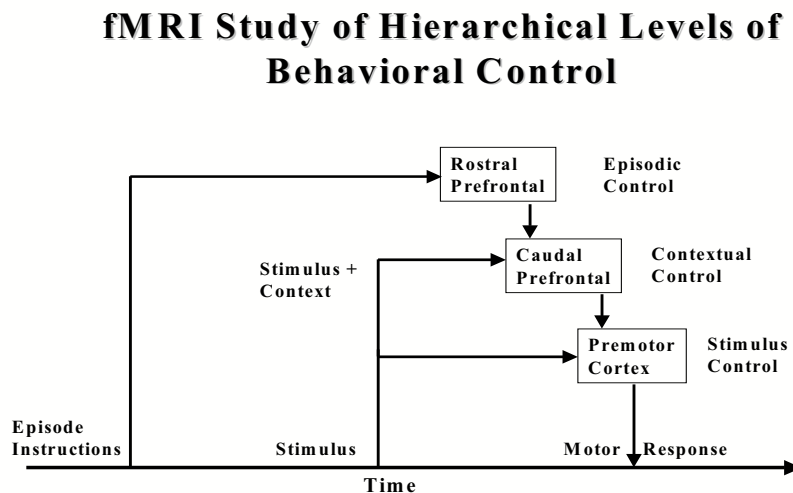
Recent research using functional magnetic resonance imaging (fMRI), reported in Science by Koechlin, Ody and Kouneiher, demonstrated a nested hierarchy of control functions within the prefrontal cortex that is very similar to MacKay's models of agency.²³ In this research, there were 3 different forms of the same task that participants were asked to do during the fMRI recording, which involved 3 different levels of sensorimotor control. At the simplest level was "sensory control" in which the appropriate response is directly signaled by the visual stimulus. "Contextual control" was established by the need to perform different tasks with respect to the same stimulus depending on the nature of an *accompanying* visual stimulus. Finally, "episodic control" was established by differences in *verbal instructions* given prior to each set of trials that set up different forms of contextual and sensory control for each set of trials.

fMRI results from this study indicate those brain areas that were most highly active during the "sensory control" task – premotor cortex, bilaterally. More frontal brain areas were activated during "contextual control" – that is, a posterior (or caudal) portion of the lateral prefrontal cortex. Finally, even more frontal brain areas were activated by the additional need for "episodic control" – that is, a more anterior (or rostral) portion of the lateral prefrontal cortex.

²³ E. Koechlin, C. Ody, and F. Kouneiher, "The Architecture of Cognitive Control in Human Prefrontal Cortex, *Science* 302 (2003): 1181-1185.

Using statistical path analysis, Koechlin and colleagues attempted to detect what structures were influencing other structures during these tasks. During “stimulus control” it was clear that the strongest interactions were between the bilateral premotor areas. During “contextual control”, influences increased between the posterior (or caudal) area of the prefrontal area and the premotor cortex. “Episodic control” activated influences between the anterior and posterior parts of the lateral prefrontal cortex. Overall, these results suggest a hierarchy of top-down control loops that can be brought into play depending on the nature of the task at hand.

Koechlin and colleagues illustrate the outcome of their research using the following flow diagram.²⁴



They argue that, with respect to the task they utilized, there are three detectable hierarchically organized forms of sensorimotor control that become activated depending

²⁴ Adapted from: Koechlin, Ody, and Kouneiher, Architecture of cognitive control in the human prefrontal cortex. *Science*, 2003, 302:1181-1185.

on task demands, and that these 3 forms of task demand involve more and more rostral aspects of the lateral prefrontal cortex:

- (1) the premotor cortex is involved in selecting *specific motor actions* depending on the current stimulus.
- (2) the posterior (or caudal) lateral prefrontal cortex controls responding with respect to the *immediate context* of current stimuli;
- (3) the anterior (or rostral) lateral prefrontal cortex keeps track of information about the *general episode of behavior* (comparing immediate and past information to the current behavioral episode).

These investigators suspect that there is a fourth, even higher-level processor in the most extreme anterior end of the frontal lobe that is involved in cognitive branching and shifting between sub-episodes of behavior, with top-down influence on the episodic control area of the rostral LPFC.

The similarity between this experimental description of a nested control hierarchy in the frontal lobes and the diagrams from MacKay is remarkable. A direct sensorimotor control loop (Sensory Control) is controlled by a meta-level feedback loop (Contextual Control), which is in turn under the control of a meta-meta-level supervisory system (Episodic Control). The point is NOT that Koechlin and colleagues have found the anatomical locations that correspond to MacKay's diagram, since MacKay did not intend his diagrams to represent anything other than minimal *functional* architecture. Rather, this research makes it clear that the sort of hierarchical control loops suggested by MacKay can actually be found in the brain; and that the system operates in a nested and top-down manner.

XI. Language

I previously described the concept of the social scaffolding of the mind. The most potent form of the social scaffolding of human cognition is language. Most of what

I have described so far regarding nonreductive mental properties point to capacities that are to some degree continuous with the capacities of animals. Language is an exception to this continuity in the emergence of mental causation. Terrence Deacon suggests that there is one discreet transition point at the human end of the phylogenetic continuum. A “symbolic threshold” has been crossed somewhere in human evolution. This threshold is crossed anew by each child via a “symbolic insight” that must be achieved during early development. Any chimpanzee that we might credit as a user of language must achieve this insight.²⁵

Lots of research in the last 20 years has suggested that the language line is not as sharp between humans and apes as we once imagined. Studies abound illustrating various forms of language-like abilities in apes. A recent article in Science even touts the large indexical vocabulary of a Border Collie.²⁶ With regard to these reports of language in non-human animals, Deacon suggests that even when the “symbolic insight” has not been achieved, there are, nevertheless, iconic and indexical communication systems clearly present in many animal species. However, Deacon rightly argues that the insight necessary for the development of fully symbolic language largely exceeds the capacity even of chimpanzees.

However one understands the uniqueness of language in humankind, human mental causation cannot be adequately discussed without consideration of the contribution of language. Unfortunately, it is not possible to do justice to the topic of human language in the remainder of this lecture. However, several general observations

²⁵ Terrence Deacon, *The Symbolic Species: The Co-evolution of Language and the Brain* (New York, W.W.Norton & Co. 1997), 73ff.

²⁶ Kaminski J, Call J, Fischer J. Word learning in a domestic dog: Evidence for "fast mapping". *Science*. 2004. 304:1605-6.

are worth making in order to link language to our current view of mental causation rooted in action loops.

First, Hutchins and Hazelhurst performed a neural network experiment that illustrates both the emergence of more and more useful symbols and the contribution of symbol systems to group problem solving.²⁷ The simulation involved a group of “citizens,” each “citizen” a connectionist neural network. Input to each citizen involved both information about events themselves, and symbols from other citizens representing these events. The task given to this community of artificial “citizens” was to predict the tides based on the phases of the moon. The simulation also involved sequential *generations* of groups of citizens that did not inherit the knowledge of the previous generation. However, new generations did inherit the symbols with their current meanings.

There were two outcomes of this simulation that are important in considering the role of language in the emergence of mental causation: (1) there was a gradual evolution of better and better symbols, and (2) the improved symbols allowed later generations to learn the environmental regularities involved in the task that earlier generations could not learn. Thus, the emergence of better symbols allowed for the emergence of better group problem solving. With respect to humans, the potentiality for solving many sorts of problems are not genetically built into the microstructure of our nervous systems, but are made available in a top-down manner from the linguistic and cultural environment to the individual.

This illustration fits nicely with Clark’s emphasis on the social scaffolding of cognition as previously described. Language symbols pre-structure thinking and problem-solving such as to allow later generations to accomplish tasks that could not be

²⁷ E. Hutchins and B. Hazelhurst. Learning in the cultural process. In C. Langton et al., eds, *Artificial Life II*. (Addison-Wesley, 1991), as reported in Clark, *Being There*, 189-190.

mastered by previous generations. Language thus provides scaffolding for both internal and external problem solving. Language is, in Andy Clark's words, "a computational transformer that allows a pattern-completing brain to tackle otherwise intractable classes of cognitive problems."²⁸ Language expands the causal potential of mental processes.

Some of the other contributions of language to cognition specifically described by Clark are: (1) offloading of memory into the socially maintained language environment; (2) use of labels as perceptually simple cues regarding a complex environment; and (3) coordination of the action of individuals and groups via internal self-talk, dialog, or in written plans, schedules, etc.²⁹

Terrence Deacon has also provided an important analysis of the role of language in the emergence of human thought in his book *The Symbolic Species*.³⁰ Some of these contributions of language emphasized by Deacon are:

- *Distancing of action from the demands of immediate motivations and needs:* Language facilitates behavior involving delayed need-gratification by augmenting the ability to consider alternative actions through entertaining various symbolic "what if" scenarios.
- *The ability to form a self-concept:* The symbolized self can become the object of evaluation. Clark also describes the importance of language in allowing for second-order cognition (or meta-cognition) in the form of self-representation and self-evaluation.
- *Expanded empathy:* We can enter into the experiences of others through emotional engagement created by stories. Such empathy also allows for the

²⁸ Clark, *Being There*, p. 194.

²⁹ Clark, *Being There*, p. 200-201.

³⁰ Deacon, *The Symbolic Species*.

development of a Theory of Mind – the ability to model and predict the mental life of other individuals.

- *A virtual common mind among groups of people*: Common semantics, metaphors, and stories create cultural groups with similar world views.
- *Ethics*: Language encodes communal values as abstract representations of “the good” that can be used for judging between potential behaviors.

XII. Imagination and Mental Modeling

Of course, there is the seeming internalness of thinking. This internalness can be understood as off-line *simulations* of action-in-the-world. The capacity for off-line symbolic thought (not expressed in bodily behavior) can be understood as *piggy-backing* on processes involved in ongoing adaptive action. Consideration of off-line versions of acting-in-the-world brings us to mind as we might understand it as existing in such internal processes as rational contemplation, conscious decision-making, and imagining the future.

A number of lines of research have demonstrated that more complex nervous systems have an ability to *simulate* action loops *off-line*. One important form of evidence comes from recent work on mirror neurons.³¹ The idea of mirror neurons comes from the demonstration that the neural systems of the supplementary motor cortex that would be involved when a monkey (in the exemplar experiment) engaged in a specific form of movement are also active in the same manner when the monkey passively observes another monkey making this same movement. Stated in human terms, the perception of the action of another person involves an implicit simulation of making a similar action our selves. To perceive action is to simulate action.

³¹ G. Rizzolatti and L. Craighero L. The mirror-neuron system. *Annual Review Neuroscience*. 2004; 27:169-92.

Another example of action simulation within the brain comes from research showing that the specific reaching behavior intended (but not executed) by a monkey can be detected in the spatiotemporal pattern of neural activity in one portion of the parietal cortex. The neural patterns also coded how badly the monkey wanted the reward that would come from the behavior (if executed).³² Thus, a simulation of the specifics of a reaching movement, as well as the desirability of the behavior, is available in the parietal cortex of the brain even when the behavior does not appear in physical action.

Researchers at Duke University went this experiment one better. They taught a monkey to control a robotic arm using a joystick, while simultaneously recording from neurons in the motor cortex. When the monkey's brain patterns for manipulating the joystick were well documented, they disengaged the joystick and allowed the brain patterns to control the robotic arm (the monkey continuing to actively manipulate the joystick). What surprised the investigators was that the monkey figured it out and spontaneously quit using the joystick, content to manipulate the robotic arm via whatever was happening in his brain. The monkey easily shifted to using his motor neurons offline to simulate, as it were, controlling the robotic arm via the joystick.³³

For many years it has been clear that simulations of the outcome of motor activity (particularly eye movements) are available to areas involved in visual perception in the form of a "corollary discharge."³⁴ Here a command to make a saccadic eye-movement is

³² S. Musallam, B. D. Corneil, B. Greger, H. Scherberger, R. A. Andersen. Cognitive Control Signals for Neural Prosthetics. *Science*, 305: 258-262, 2004.

³³ This experiment was described by Miguel Nicolelis of Duke University on CBS News, 60 Minutes, and can be found at: www.cbsnes.com/stories/2003/10/13/tech/main577757.shtml. Also see Carmena JM, Lebedev MA, Crist RE, O'Doherty JE, Santucci DM, Dimitrov DF, Patil PG, Henriquez CS, Nicolelis MA. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*. 2003;1(2):E42.

³⁴ The phenomenon of corollary discharge is described in B. Kolb and I.Q. Whishaw. *Fundamentals of Human Neuropsychology, Fifth Edition*. (New York: Worth Publishers, 2003), 403-4044.

communicated both to the lower brain nuclei controlling eye-movements and to the visual cortex. The signal to the visual cortex is the corollary discharge. Based on this information, the visual input created by the eye movement is not interpreted by the visual system as a movement of the world. The visual world appears stationary as we move our point of eye fixation from place to place. The corollary discharge is presumed to be some form of pre-action simulation of the sensory impact of the eye movement.

To shift to a higher-level example of behavioral simulations, we turn to the early work of the psychologist Wolfgang Köhler. Köhler's work on insight in apes suggests that *problem-solving solutions* can be tried out in the form of internal mental simulations prior to behavioral execution.³⁵ Köhler describes the attempts of a chimpanzee named Sultan to obtain some bananas outside his cage. Some sticks were available, but none of them were long enough for Sultan to reach the bananas. After many unsuccessful attempts at solving the problem, Sultan quit trying and sat in the corner, apparently ruminating over the problem for a while. Then, he suddenly went back to the problem and immediately executed the proper solution by putting two sticks together end-to-end allowing him to reach the bananas. It appears as though Sultan was able to solve the problem via off-line, imaginative, mental modeling prior to executing the correct solution. Of course, this experience of off-line problem solving is common to human beings, but an important point of this illustration is that such off-line simulation of behavioral scenarios occurs in the absence of language and self-talk.

Given these examples, it appears that the brain is adept at sensorimotor simulation. It is clear from the evidence described that the brain is adept at running off-line simulations of action loops, whether simulations of simple motor actions or complex

³⁵ Köhler, W. *The Mentality of Apes* (New York: Harcourt, Brace, 1927).

problem-solving. Action-feedback-evaluation-action loops can be run in off-line simulations in order to predict the likely consequences of prospective action.

When combined with the ability to use language, off-line behavioral scenarios become a form of inner problem solving using self-talk. A number of cognitive psychologists believe that all thinking is done with inner speech. Subjectively, inner speech seems to be the predominant mode of conscious thought. It is, however, not likely to be the only form of off-line mental modeling available to humans. The main point, however, is that internal, imaginative problem-solving and thinking piggy-back on action-loops via the ability to run sophisticated sensorimotor simulations off-line.

XIII. Consciousness

Finally, we need to consider the role of consciousness in behavioral flexibility and adaptability in order to complete this physicalist account of mental causation. The role of consciousness and subjective awareness can be readily understood by consideration of the disabilities and residual abilities of individuals with disturbances of consciousness due to certain forms of brain damage.

Much has been written in philosophy of mind about the phenomenon of “blind sight.” This phenomenon occurs in persons with damage to the visual cortex on one side of the brain who consequently lose the ability to see anything in the opposite side of the visual world (i.e., there is a one-sided loss of conscious visual perception). “Blind sight” refers to the ability of some of these individuals to reach out and intercept a target that is moving within the area of blindness that they nevertheless report that they cannot see. Some would suggest that this proves that phenomenal awareness is unnecessary for action. The correct interpretation is that phenomenal awareness is unnecessary to some primitive forms of action (intercepting a moving target), but necessary for other forms of action. Thus, when the moving target moves from the area of blindness into the

area of visual perception and awareness, the person can consciously and verbally identify the moving object. For example, is the object a fly or a wasp – a Frisbee or a football? Phenomenal awareness of stimulus meaning allows the person to consider different possibilities for action with respect to the flying object based on conscious perception of the nature of the object, and to verbally name or describe the object. Thus, phenomenal awareness opens the possibilities for a wider and more flexible range of adaptive actions.

The contribution of consciousness to behavior is also apparent in the contrast between individuals with anosognosia versus individuals with other forms of agnosia. Anosognosia is the neuropsychological term used to identify a disorder (usually from damage to the right parietal lobe) that involves paralysis of one side of the body, with *adamant denial* of the disability. Somehow the body representation of these individuals cannot be updated in such a way as to provide information to conscious awareness regarding their unilateral paralysis. Thus, they are not able to adjust their behavior to take the paralysis into account. In contrast, other forms of agnosia (“agnosia” meaning “absence of knowledge of”), such as the inability to recognize faces, create deficits regarding which the patient is acutely aware. Presence of phenomenal awareness of the disability allows patients to adjust their behavior to accommodate the deficit.

Finally, the fractioning of the content of consciousness created by severing of the connections between the right and left cerebral hemispheres (resulting in individuals with a “split-brain”) is also revealing of the nature of consciousness. The surgical procedure (done for the relief of otherwise intractable epilepsy) severs all neural connections between the cerebral cortices, including the 200 million axons of the corpus callosum. The result of this procedure (repeatedly demonstrated in a variety of experiments) is that sensory information, cognitive processing, and motor control within each hemisphere are isolated from processing in the other hemisphere. For example, information occurring

only in the left side of the patient's visual world (the snow scene in left part of the viewing screen in this illustration) will be seen only by the right hemisphere and, therefore, can be responded to only by the patient's left hand (controlled by the right hemisphere that is privy to the visual information). However, the right hemisphere of the split-brain patient will have no awareness of the image of the chicken claw. What has been suggested is that the split-brain patient now has split consciousness. Each hemisphere is conscious of different information without the benefit of the sharing of information from the opposite hemisphere. The consequences of cutting the corpus callosum are different contents of consciousness in each hemisphere, and different outcomes in behavior for the two hands. Since this split-brain procedure had created two separate domains of consciousness, consciousness itself is a property of the physical operation of the brain.

We can see that consciousness is an important in allowing for greater flexibility and adaptability of human behavior. Although sugar-factory experiment (described earlier) suggests that we should not reserve the idea of mental causation exclusively to that which is conscious, conscious awareness nevertheless significantly increases the range of possibilities for the modulation of ongoing action, that is, for mental causes of action.

XIV. Summary

This paper has attempted to enrich our imaginations for the possibilities of physicalism by outlining the bare biological bones of the emergence of efficacious mental function. In simple outline, the emergence of high levels of mental causation and moral agency has come about by development of brain complexity allowing for:

- (1) more and more hierarchically nested action-evaluation-and-modification loops;

- (2) the capacity to run action-loops off-line in sensorimotor simulations of potential behavior;
- (3) symbolic representation (language) and the enormous amount of external scaffolding inherent in human culture;
- (4) an increasing role of consciousness in the modulation of behavior.

These capacities are elaborations of a basic biological process involved in adaptive action-loops. However, from these basic processes have emerged increasingly efficacious forms of mental causation.

Moral agency comes about as one learns procedurally and verbally what is good, beneficial, socially acceptable, and Biblically mandated or encouraged. These behavioral procedures and symbolic concepts become critical evaluative criteria for both on-line and off-line modulations of thought and behavior, embodied within hierarchically nested action loops.

In this final lecture, I have been motivated by the hope that the examination of these bare bones might reduce some of the mysteries and enhance our imaginations regarding the emergence of embodied and embedded mental causation and moral agency. I believe these biological bones provide a skeletal framework for a physicalist and non-reductive understanding of complex and robust forms of human mental causation, including those forms of mental causation necessary for presuming human beings to be responsible moral agents.

The answer to the question of the title of this lecture – “Did my neurons make me do it?” – is “No, I did it. My neurons are merely a part of the integrated and highly complex physical ME.” My brain systems and their neurons are a necessary part of my doing, and their absence or dysfunction would limit my capacities to regulate and evaluate my actions. However, attribution of responsibility for my actions to any subpart of me is to ignore the nonreductive nature of human physical embodiment.

Bibliography

- Anderson, J.R. 2000. *Cognitive Psychology and Its Implications*. New York: Worth Publishers.
- Bargh, J.A. and Chartrand, T.L. 1999, The Unbearable Automaticity of Being. *American Psychologist*. 54, 462-479.
- Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, Mass.: Bradford Books.
- Deacon, T.W. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain* New York, W.W.Norton & Co.
- Deacon, T.W. 2003. The hierarchical logic of emergence: Untangling the interdependence of evolution and self-organization. In *Evolution and Learning: The Baldwin Effect Reconsidered* edited by B. Weber and D. Depew. Cambridge: MIT Press.
- Deacon, T.W. 2001. Paper at the Conference on *Science and the Spiritual Quest*, Boston.
- Dretske, F. 1998. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, Mass.: Bradford Books.
- Heisenberg, M. 1994. Volunarity (Willkürfähigkeit) and the General Organization of Behavior. In *Flexibility and Constraint in Behavioral Systems*, edited by R.J.Greenspan and C.P.Kyriacou. John Wiley & Sons, Ltd.
- Hutchins, E. and Hazelhurst, B. 1991. Learning in the cultural process. *Artificial Life II*, edited by C. Langton et al., eds. Addison-Wesley, 1991.
- Kolb, B. and Wishaw, I.Q. 2003. *Fundamentals of Human Neuropsychology, Fifth Edition*. New York: Worth Publishers.
- Mackay, D.M. 1991. *Behind the Eye*. Oxford, UK. Basil Blackwell Ltd.
- Murphy, N. 1999. Supervenience and the downward efficacy of the mental: A nonreductive physicalist account of human action. In *Neuroscience and the Person: Scientific Perspectives on Divine Action* edited by R.J. Russell, N. Murphy, T.C. Meyering, and M.A. Arbid, Vatican City State: Vatican Observatory.