

Neuroscience, the Person and God:

An Emergentist Account

by Philip Clayton

Abstract. Strong forms of dualism and eliminative materialism block any significant dialogue between the neurosciences and theology. The present article¹ thus challenges the “Sufficiency Thesis,” according to which neuroscientific explanations will finally be sufficient to fully explain human behavior. It then explores the various ways in which neuroscientific results and theological interpretations contribute to an overall theory of the person. Supervenience theories, which hold that mental events are dependent on their physical substrata but not reducible to them, are explained. Challenging the determinism of “strong” supervenience, I defend a version of “soft” supervenience that allows for genuine mental causation. This view gives rise in turn to an *emergentist* theory of the person. Still, I remain a *monist*: there are many types of properties encountered in the world, although it is only the one nature which bears all these properties. The resulting position, *emergentist monism*, allows for diversity within the context of the one world. This view is open at the top for theological applications and interpretations, while retaining the close link to neuroscientific study and its results. Theology offers an interpretation of the whole world based on a yet higher order of emergence, although the notion of God moves beyond the natural order as a whole. It therefore supplements the natural scientific study of the world without negating it.

Acknowledgement

This article is adapted and reprinted from *Neuroscience and the Person: Scientific Perspectives on Divine Action*, ed. Robert John Russell, Nancey Murphy, Theo C. Meyering, and Michael Arbib, © 1999 Vatican Observatory Publications and the Center for Theology and the Natural Sciences.

1.1 What Theologians Seem to Want

Much could be (and has been) said about neuroscience from the perspective of theology; theologians (in this book and elsewhere) have not been shy about offering their own interpretations of brains, thoughts and spirits. Of course, turnabout is fair play: neuroscientists have also not been shy in commenting on theology, God, free will, and other concepts both human and divine. After all, it is in part an empirical question why a brain with a structure like ours—one that shares our evolutionary history, stores and retrieves information as ours does, and responds like ours to electrochemical stimulation and viruses—would produce religious ideas and religious experience as ours does.

As fascinating as such disputes might be, they are not the subject of this paper, at least not directly. For I am convinced that direct battles between neuroscience and theology (or, for that

1. In writing this piece I have accrued more debts than I can possibly pay here. Each of the members of the Vatican/CTNS neurosciences team directly or indirectly influenced what appears here. Among them, Arthur Peacocke and Thomas Tracy must be singled out for supportive comments and helpful criticism (respectively). In their role as editors of (Russell et al. 1999) Theo Meyering and Nancey Murphy wrote detailed commentaries and criticisms, which resulted in significant improvements. That imperfections still remain reflects more on the author than on the critics.

matter, direct concordances) will not even become conceivable until a new, deeper mediation between the two fields has been achieved. The neurosciences raise a question much closer to home than disputes about God: the question of who *we* are. Progress in neuroscience challenges, or at least is often taken to challenge, cherished notions of what it is to be a human person: self-consciousness, soul, “thinking being,” free will. Unless and until we manage to defend a notion of the person that preserves concepts such as these *in light of* what we now know about the human brain, language about God, and any work such language is supposed to do within the human mental psyche, will appear gratuitous.

This is not to say that theological doctrines cannot be of any assistance.² Some essays in this book describe theological doctrines that impinge, directly or indirectly, on work in the neurosciences and the social sciences. Imagine, for example, how one’s notion of personhood might be affected by acceptance of the Christian doctrine of creation, the belief that we are “dust” and yet nonetheless indwelt by the Spirit of God. Divine creation would introduce purpose, for example, and purpose means an arrow of time—the belief that, besides the brute-fact given of the natural world, there is the directing influence of an unconditioned will who is both source and telos of this world. Christian theologians then supplement creation beliefs with content from the biblical texts (and their tradition(s) of interpretation), which are taken to provide an important record of divine communicative intent. When the theologian has this much, she already has huge constraints on her theory of personhood. She has, implicitly, a Christian ethic of “willing the will of God.” Moreover, the biblical texts speak of *covenant*, which implies a set of divine commandments for living—but also mutual agreement and responsibility (hence ethics again). Covenant gives rise to the notion of sin, which Thomas Tracy defines elsewhere as “the bias of the will toward an orientation of alienation from God” (Tracy 1999). The idea of sin then gives rise to the idea, and thus the possibility, of reorientation. Possibility is taken to have become actuality through divine initiative or grace.

I include this well-known list to show how detailed is the anthropology (theory of human nature) that theologians bring to the discussion table. (*Nota bene*: “detailed” does not entail “non-negotiable.”) It includes, at least for traditional theologians, not only the existence of at least one purely spiritual being—hence the possibility of disembodied agency—but also the notions of will and of freedom, which come in both finite and infinite flavors. With will, so understood, comes consciousness: Christians conceive God as conscious agent, an agent enough like human agents that the predicate “person” can also be attributed, if only in an analogous fashion, to the divine. On this view, humans and God are also *moral* agents; persons exercise their agency in light of real obligations to other persons (indeed, to the world as a whole) and to God. Finally, these agents are *social* agents. Religious notions of *community* emphasize a union among humans in light of the divine presence and the covenant which makes of us “one.” The christological expression of community is *kerygma*, the particular story of Jesus Christ, culminating in the belief in an eschaton or

2. I am grateful to Mark Richardson for discussions of the following theological constraints on the theory of personhood.

second coming. In short, the Christian parameters for talking about persons stretch from the moment of creation to the consummation of history, and from individual birth through life and death and on to the hope of a final reconciliation.

1.2 What Neuroscientists Know

How does theological anthropology contrast with recent results in the neurosciences? To get a flavor for the sort of data and emphases, consider just a few of the recent findings on the physiological bases of human cognition, which offer a representative sample of the sorts of connections that have now been established:

* Recent advances in genetic research have led to a better understanding of how sensory receptors function to convey sensory inputs from the environment to the brain. By comparing the genetic constitution of the receptors (e.g., receptors for smell and taste) from many different species, we are beginning to understand how sensory receptors first evolved and how they subsequently became more highly specialized in higher primates and eventually in *Homo sapiens* (Hodos and Butler 1997).

* Studies involving the prefrontal cortex of humans and monkeys have led to a deeper understanding of the dynamics of short-term memory. Using implanted electrodes and PET scanning, specific areas of the prefrontal cortex have been demonstrated to be involved in spatial working memory, performance-ordered tasks, verbal working memory, object working memory, and analytic reasoning. These measurement techniques allow neuroscientists to study subjects as they perform a variety of tasks and to determine which specific areas of the brain are being stimulated during which activities (Beardsley 1997).

* Studies of patients with frontotemporal degeneration (FD) and patients with Alzheimer's disease (AD) have led to further insight into how the brain functions in the use of language. Persons suffering from FD show damage to the left frontal and anterior temporal brain regions, which leads to an accompanying decline in grammatical abilities. AD patients have defects from cerebral profusions in the inferior parietal and superior temporal regions of the left hemisphere, which correlates with a decline in semantic ability. A recent study by Grossman et al. demonstrated that, although language ability is centered in the left hemisphere, no single brain region is responsible. Instead, linguistic ability is the product of "a neural network distributed throughout the left hemisphere subserving different aspects of language comprehension" (Grossman et al. 1998).

* People suffering from Huntington's Disease often show frontal lobe atrophy (and eventually total brain atrophy). Early manifestations of frontal lobe atrophy include specific cognitive losses: problems with attention, concentration, planning, and memory. Similar cognitive losses are characteristic of persons suffering from lesions in the frontal lobe (Aylward et al. 1998).

* Recent studies have demonstrated that William's Syndrome is a result of the loss of the end of one of the copies of chromosome 7, involving perhaps 15 or more genes. People who suffer from this condition are typically diagnosed mildly to moderately retarded, scoring in the 50s to 60s on IQ tests (equivalent to people with Down's Syndrome.) However, while people suffering from William's Syndrome have limited writing and arithmetic abilities, their verbal communication skills and their

ability to recognize faces and facial features often surpass those of their non-William's peers. In addition, some WS patients have an almost uncanny musical ability, and close to perfect pitch, even though they are unable to read written music. Although their overall brain mass is reduced, the frontal lobes are preserved (including the temporal lobes, which are associated with visual memory); one also finds an enlarged primary auditory complex and a comparable limbic area, which is associated with emotion (Lenhoff et al. 1997).

* Corticospinal excitability has been studied using focal, single-pulse transcranial magnetic stimulation (TMS) applied to the scalp. Using fourteen right-handed subjects, the researchers studied the effects of TMS on motor ability on each side of the body. When the subjects induced themselves to think sad thoughts, this facilitated greater motor potentials in the left hemisphere of the brain, while self-induced happy thoughts evoked greater motor potential in the right hemisphere. Tormos et al. take these results as a clear sign that the brain evidences some form of lateralized control of moods (Tormos et al. 1997).

Results such as these present a clear challenge to those who would rend thought and affect from its physical substratum. The influences are both deep and bi-directional; they involve the deepest areas of mental functioning. Whether humanists and religionists are ready to acknowledge it or not, the neurosciences are now producing alternative explanatory candidates in the study of the human person.

1.3 Putting the Pieces Together

In this paper I will argue that the perceived tensions between theology and the neurosciences call for *renewed reflection on the nature of the person*. Formulating a philosophically adequate account of what it is to be a human person provides crucial guidance on how to relate these two diverse fields. Conversely, attempts at a direct connection (or reduction) of the one area to the other threaten to mislead unless one keeps the question of human personhood at the center of attention.

One feature of this discussion should be clearly acknowledged from the outset: in the nature of the case, physical sciences such as the neurosciences do tend to push one in the direction of *physicalism*, the view that all things that exist are physical. For it is a basic assumption of good neuroscience, as with the other natural sciences, that only traceable physical causes be employed and that only physical mechanisms be introduced in explanations. "Physical" here has an interesting double meaning: it is a methodological standard (physical means "the sorts of things that physicists can study") as well as an ontological thesis (something is physical if it is built up out of the fundamental particles and energies that we have discovered in the natural world).³ So the question

3. To say that physics is committed to *materialism*, the view that all things that exist are composed out of matter (or basic material particles such as atoms) would be to assert that physicists are committed to a particular metaphysical thesis; but this is to place metaphysics prior to the actual conclusions and methods of science, a move we should reject. What is interesting about *physicalism* is that, if it's to have any special warrant, it must appeal to the actual results of physics: the physicalist is committed to the ultimacy of just those entities and explanations that our best physics commits us to (or: the sorts of entities that physics could eventually come to know). Given this definition, to resist physicalism, as I do, is to resist the claim that all

cannot be, Should we do a different kind of neuroscience, say, one that talks more about souls and their actions? Instead, the guiding question in the dialogue between theology and the neurosciences is, How far can a position go in the direction of the physicalist assumptions that are basic to the empirical study of the brain without denying (or implicitly rejecting) factors necessary for the viability of religious belief?

To start with the obvious: holding a physicalist view of the world that denies the very existence of God—perhaps on the grounds that God is not a physical being, hence (given physicalism) God cannot exist—would leave little or nothing over of traditional Christian metaphysics. Short of ruling out the existence of God, however, one might find oneself offering interpretations of neuroscience that fundamentally conflict with belief in God or in divine action—which would represent virtually as complete a rejection of Christian theology as an outright denial of the existence of God. Or (somewhat less obvious but still crucial) one might be tempted to adopt a theory of knowledge that makes all theological accounts extremely unlikely (or meaningless) as explanations (cf. Clayton 1997). Finally, there are views and approaches that make the concept of God explanatorily unnecessary, a sort of appendix on the overall body of explanation. Although the last three views are not immediately fatal to Christian belief, they cast this belief deeply into question, opening the door for a well-argued case that “we have no need of that hypothesis.”

I suggest that *these* sorts of tensions, rather than direct disputes over theological claims, actually represent the core of the discussion with the neurosciences. Before one turns to concrete theological proposals, one needs to be fully aware of the stakes of the debate. If one holds that theological assertions are mere constructs or “evocative metaphors,” then she has already ceded the case; she might as well just admit up front that theology does no explanatory work whatsoever. If theology is to *do* anything in this discussion, then it must first be shown that there is something about the human person which *cannot* be captured, directly or derivatively, in neuroscientific terms.

2. Methodological Parameters for the Neuroscience/Theology Debate

2.1 *Some Views on the (Actual or Potential) Impact of Neuroscience on Theology*

Before entering the actual debate, consider a few of the now-dominant views on the probable impact of neuroscience on theology. In what follows I will challenge the first five and will defend a version of the sixth.

(a) One position, which we might call *the Arbib Credo*,⁴ states that all data about the human person will eventually be best explained in neuroscientific terms. On this view, theology has no

adequate explanations will ultimately be given in terms of physical laws and entities (or in terms of some idealized version of present-day physics).

4. I use this title in deference to Michael Arbib, who has presented one of the most sophisticated alternatives to theism; see, among many other works, (Arbib 1999) and the literature cited there.

explanatory power of its own; its terms pick out no theological entities or properties in the world but are rather constructs of individuals or societies. Moreover, says the Credo, the neurosciences (among other sources) give us sufficient reason to abandon the traditional explanatory claims of theology.

A variant on this view, which one might call “*watch-out-ism*,” admits that nothing in current neuroscience falsifies theology ... yet. But it warns that the results of neuroscience will eventually be disastrous for theology. For example, she argues, we will eventually be able to predict an individual’s actions based on the inputs from his environment and the succession of brain states, at which time free will will be in trouble. Or, she argues, we’ll eventually understand the neural factors that incline people to have what they call “religious experience,” at which time religious beliefs will cease to serve as credible explanations of religious experience

(b) The opposite position relies on *soul-based explanations*. On this view, souls exist. They are the kind of things which can be explained only with theological terms such as “the image of God,” salvation or immortality. The nature of the soul makes it inaccessible, even in principle, to neuroscientific investigation.⁵

(c) One might also hold a view of *instrumentalism and agnosticism*. On this view, the neurosciences help us to understand human behavior and cognitive functioning; they are indispensable because of their usefulness in prediction and scientific understanding. But this view is agnostic about theological questions; it views them as simply falling into another category altogether, one not addressed by neuroscience. .

(d) The *No Conflict* view believes it can overcome the potential conflicts presupposed in (a) and (c). It does so, however, by making all the changes on the theological side: theology means something different than one used to think. One way to avoid any conflict is to espouse a purely naturalist theology, one which doesn’t claim the existence of any super-natural beings or properties. Theologians might still speak of the spiritual significance of the universe, using labels such as “sacred” or “ecstatic” naturalism, but on this view they do not (should not) mean these terms in a way that conflicts with physical explanations of all phenomena in the world. A related way to remove any possibility of direct conflict is to view all theological statements as metaphors in such a way that any actual or possible conflict is ruled out *ab initio*.

(e) Let us call the view *compatibilism* which holds that the concrete results of neuroscience neither prove theology nor disprove it. Rather, on this view, the data are at least consistent with what one would expect to find if, say, a Christian theological anthropology is true. For example, theology also might teach a theory of human nature in which conscious life is biologically-based. Of course, consonances of this type do not prove the truth of Christianity, but they do tell against “conflict” views of the neuroscience/theology relationship. In this paper I defend *emergentist monism* as one

5. Of course, as Stephen Happel has pointed out, “soul” could be, and has been, used in less dualist senses, e.g. to emphasize the religious dimension in psychological and even biological phenomena. This is surely correct. Because soul-language has traditionally served as the bulwark of dualist metaphysics, however, I use it here as the antithesis to the Arbib Credo and avoid it in my own constructive proposal below.

viable compatibilist answer.⁶

2.2 *Four Models of the Rationality of Religious Belief*

I have argued elsewhere that scientific results are rarely the direct building blocks of theology: seldom are theological explanations constructed, directly or indirectly, using the language of scientific results and explanations (Clayton 1989; Clayton 1998a). Instead, the results of science must first be metaphysically interpreted and their underlying assumptions brought to the surface before they can tell for or against theological assertions.⁷ Such meta-physical or meta-scientific assumptions have to be made explicit—and sometimes fought over—before one can get to the really interesting discussions. (In fact, most of the underlying assumptions of science are *not* directly theological in nature, even though we focus here on those metaphysical assumptions that can help to structure the debate between the empirical sciences and theology.) If the theology/neuroscience discussion turns in large part on clarifying the framework assumptions (such as a theory of personhood) that determine how the two fields are to be connected, then we must come to agreement on what standards of evidence to apply to meta-empirical theories of this sort. I suggest four possibilities:

(a) One might hold to the standard of *proof*, requiring that a compelling (or even logical) demonstration be given of any meta-empirical or religious belief. This was the sort of approach favored by natural theologians earlier in the modern period. The difficulty with such standards is that rigorous deductive proofs don't seem to be available even in most of the natural sciences (e.g., in most of the biological sciences). Further, linear proofs represent a standard inappropriate to any hermeneutical discipline, that is, to any discipline whose structure involves a two-way movement or interdependence between data and explanatory theories (Clayton 1989, chap.3).⁸ Since religious questions are clearly hermeneutical in this sense, the standard of proof seems unavailable from the start.

(b) One might then only require that religious beliefs be *empirically probable*. Beliefs are empirically probable when they are based or grounded on empirical evidence. This is the standard that is represented by the natural sciences, although philosophers of science have raised serious

0. "Compatibilist" in this sense must not be confused with compatibilism in the free will debate, which is the view that moral responsibility is compatible with all of one's actions being genetically or environmentally determined. Note also that compatibilism in this sense is weaker than an entailment relationship; it does not imply that results from neuroscience will someday be able to decide between (say) a Christian and a Muslim anthropology.

7. Theo Meyering correctly points out another type of science/theology connection: "meta-inductive or speculative *extrapolations* that may be more or less plausibly drawn from the respective theories and thus are at best associated with them rather than being implied by them."

8. Clearly the natural sciences are hermeneutical in the sense that they raise questions of interpretation (and are pursued by interpreting subjects). But there is an added interpretive dimension, which Anthony Giddens calls a "double hermeneutic," in the human (and theological) sciences: interpreting Shakespeare (or salvation history) involves intentional agents both in the act of interpretation and in creating the object to be interpreted.

skeptical objections of late about the claim that one can move directly from empirical evidence to theoretical explanations. Of course, there are obvious problems with demonstrating the empirical probability of religious belief, since (unlike science) the beliefs in question involve truths and being(s) that are claimed to transcend the empirical universe. The claim that there can be knowledge only of empirical matters, never of metaphysical ones, traces back to Immanuel Kant, although there are now reasons to be suspicious of Kant's sharp dichotomy (see Clayton 2000, chap. 5).

(c) A weaker standard would be to say that religious beliefs need only *not be counterindicated* by the empirical evidence. According to this view direct empirical probability is not required; still, one's beliefs should be in no worse shape in the face of the available evidence than are the competing beliefs. If one is aware of evidence against one's position, for example, one has the obligation to reject that position.⁹ I agree that theologians ought to submit themselves to this standard.

(d) A final position holds that empirical evidence is simply *irrelevant* to religious belief. This view, often called *fideism*, holds in its most extreme form that faith itself is sufficient to ground belief. But I shall use the term in a looser sense to stand for any position which does not think that even countervailing empirical evidence is sufficient reason to reject a religious belief or set of beliefs.

My own preference is for (c). This is not special pleading, for there is good reason, on the one hand, to hold that many of our deeply held beliefs, religious and otherwise—beliefs about which it seems extremely counterintuitive to say that one *ought not* to hold them—do not meet the standards of proof or empirical-scientific probability. On the other hand, agents are justified in rejecting a position when they themselves have reasons to think that it is false, and this pushes one toward (c) over (d) (see Clayton and Knapp 1994; Richardson and Wildman 1996, Part II).

What is the significance of siding with (c)? It implies that one's metaphysical beliefs are not and do not need to be direct inductive inferences from the empirical world as it's known through the natural sciences. Hence (c) means siding with fallibilism, in the sense advocated by philosophers of science such as Karl Popper and Imre Lakatos, rather than with positivism or other induction-based theories of knowledge. Each person holds a variety of meta-empirical beliefs; she may well be justified in holding these beliefs, even if they are not directly grounded in experience, as long as she holds them in a nondogmatic manner. A fallibilist epistemology means, first, that one will look for *conflicts* with experience—say, possible conflicts with the results of science—and that one will change her higher-order beliefs appropriately. It means, second, that one will observe carefully whether her various beliefs about the human person (religious, empirical, experiential) fit together into a theory that is both internally coherent and scientifically viable.

9. This epistemic standard draws significantly from the theory of falsification in the philosophy of science, as developed by Karl Popper and modified by Imre Lakatos. For further references, see (Clayton 1989) and (Murphy 1990). There are also parallels to the theory of defeaters and "defeater defeaters" developed by Alvin Plantinga, e.g. (Plantinga et al. 1983).

2.3 *The Insufficiency Thesis*

The debate about neuroscience, psychology and mind presents one with a confusing clutter of possibilities. And yet in one sense one finds herself returning again and again to one basic choice. Many neuroscientists, but not all, maintain *the Sufficiency Thesis*. It is the view that in the future neuroscience will be sufficient to explain all that we know about the human person. This view contrasts with *the Insufficiency Thesis*, which predicts that neuroscience will *not* be sufficient to explain all we come to know about the human person. I defend *the Insufficiency Thesis* in what follows not because of blindness to the power of the neurosciences (far from it!), but because there are parts of what it is to be a person that lie in principle beyond their reach. This “something more” has been called variously *consciousness* (Chalmers, Nagel, Jackson, McGinn), *original intentionality* (Searle), or perhaps *caring* (Haugeland). What it is and why it should play this role will concern us further below.

Note some of the major features of the debate between the Sufficiency and Insufficiency Theses: (a) It is not settled by any current empirical data. (b) It is future-oriented. Indeed, its status is closest to that of a *wager*. Current scientific results and scientific progress to date are relevant to which side I wager on and how much I am willing to wager (i.e., how strong is my commitment to the one view or the other). But other assumptions—metaphysical assumptions—also play a role in the different predictions of the Sufficiency and Insufficiency theorists. Think, for example, of the stock market: those who invest are willing to wager money on a future state which no one knows for sure. Even when one is a specialist in the vast amount of data that indicate whether one should invest in one or another firm, every investment remains a speculation.

(c) *Sufficiency vs. insufficiency* is, in this sense, a classic philosophical debate, not itself a scientific debate. In this sense it is more like the debate about universals or free will than like the question of the explanation of thermodynamic phenomena. For example, among the Greeks Democritus might have been on the side of the Arbib Credo and Aristotle on mine.

(d) The debate bypasses the debate about dualism. Like “positivism,” the word “dualism” seems today to be used only as a term of derision, at least in debates with or written for neuroscientists. Dualism is, strictly speaking, a species within the genus *substance ontology*, that is, it is a theory of being in which the world is divided into two basic types of existing things called substances. As such, it presupposes that a theory of being can and ought to be developed, an assumption not made in this paper. But the real differences and interesting questions raised by the neurosciences today are not adequately grasped within the framework of (traditional) substance ontology. So defending the Insufficiency Thesis is not the same as advocating dualism.

(e) The Insufficiency Thesis is compatible with believing in the great explanatory power of neuroscience. It need not be an *anti*-scientific position, and I do not advance it as such. Nothing in the present paper blocks or diminishes the importance of neuroscientific research. It denies only one thing: the final sufficiency of the neurosciences for explaining the human person.

3. Toward a More Productive Debate on Neuroscience and Personhood

3.1 Progress in Neuroscience

From the outset one should be honest about how strong the tug is in the direction of the *Arbib Credo*; there is no point in hiding one's head in the sands of a pre-scientific age that denied the dependence of the mental on the physical. As we saw above, specific types of cognition—perceptions, memories, emotions—do correlate with specific state changes in specific brain regions. In some cases we can predict with a high degree of accuracy what neural processes will accompany which sorts of subjective experiences. Note that knowledge of the connections is increasing in both directions: neuroscience can predict more of the subjective experiences that will follow specific types of brain stimulation, as well as more about the sorts of neural activity (and in what regions) that will underlie particular psychological experiences.

Specific types of brain damage also lead to specific changes in subjective experience. In one famous case, Weiskrantz reports on a subject whom he calls “D. B.” who had had part of his visual cortex removed so that he was unable to see things located in a certain part of his visual field (Weiskrantz 1987). But if D. B. was asked to *guess* what the supposedly invisible object was, he could do so with nearly 100% accuracy. This phenomenon has come to be known as “blindsight.” Weiskrantz suggests that D. B. may have been drawing on information located in the lower temporal lobe, suggesting that there are regions that are necessary for *conscious* awareness of what one perceives, even though one may draw information from other regions of whose processing he has no conscious awareness. (Interestingly, D. B. was able to learn with some practice to have at least limited awareness of what he “knew” in the lower temporal lobe.)

The brain sciences have thus established that, and how, specific types of mental experience correlate with specific brain functions. It is the *brain* that does the processing when you calculate 73×37 or when you feel fear after hearing a threat. In coming years we will learn massively more about what neurological states are *necessary* for certain mental experiences; we will find more and more such necessary conditions; and we will be able to specify the underlying brain states and processes with greater and greater precision. Gradually, we will be able to cause more and more specific mental or emotional responses by means of carefully controlled stimulation to the brain, and we will be able to model more and more of them on computer-based systems.¹⁰

As the neurosciences develop, we will be able to give increasingly complete accounts of how perceptions are represented, how they are recalled, and what is happening in the brain when a subject reports that one thought gives rise to another. We'll understand the functions of emotions and why brains that have emotions like ours would confer survival value on an individual. We'll also learn

10. I recall reading recently in the popular press that Air Traffic Control already has an awareness meter that allows supervisors to monitor when an air traffic controller is losing conscious attention (“dozing off”). Perhaps it would be useful for professors to employ awareness meters for those students who tend to doze off during lectures!

precisely which brain regions or distributed systems are active when a person reports having certain emotional, aesthetic or religious experiences. We'll know why brains such as ours would be prone to aesthetic and religious experiences of these sorts and what kinds of neural stimulations (or lesional damages) tend to increase or suppress such experiences. Some argue—though others dispute—that in the limit case we could learn the precise brain states that would have to occur if a human subject is to enjoy particular kinds of mental (phenomenal) experience.

Now some readers may find the prospect of such successes in neuroscience greatly exciting; others may find it greatly threatening. Whether a massively successful neuroscience would be a good or bad thing is not the topic of the present paper. Instead, I want to ask: if neuroscience is successful in this way, will we have *proven* that all things that exist are physical, hence that the conscious self is an illusion? By no means! To move from successes in neuroscience to the doctrine that only physical things exist—whether one then advocates the falseness of belief in God or interprets the self as “merely metaphorical or constructed”—is, as I shall attempt to show, a category mistake. First let's look at some of the options, and then I shall make a case for a mediating framework that I think is preferable to the two extremes.

3.2 Getting Rid of the Extremes

The sorts of neuroscientific results that I have just summarized (or at least imagined) tend to pull people in one of two directions. Some find here strong evidence that human cognitive behavior will ultimately be fully explained in terms of brain activity (the Sufficiency Thesis). Others find here no threat to their dualist intuitions: thought and emotion are still properties of the “spiritual self,” they insist, and spirits or souls are just not the *kind* of thing that brain science can really tell us anything about. I will argue that *neither* of these more extreme views does justice to the data we have about the human self. There are serious issues in neuroscience and religion, but they depend on drawing careful distinctions closer to the middle and not on battles fought at the edges.

Let's call the two more extreme positions I've just mentioned *strong reductionism* and *metaphysical dualism*. Strong reductionists argue that human thought and mentality is in principle fully explainable by, because wholly caused by, neural firings. To understand why regions of the brain react in the ways they do would be to understand human thought, human emotions, human religious experience. According to so-called identity theorists, thoughts *just are* the neurological events studied by brain scientists (Armstrong 1993; Churchland et al. 1992; Churchland 1986).

Metaphysical dualists on the other end of the spectrum argue that there is an ontological entity such as the soul or mind (*Geist*) that is forever inaccessible to natural scientific study, even in principle, and that is the basis for and possessor of all mental events: ideas, wishes, emotions, intentions, and the like.

Other authors in this book have given good summaries of why dualism is no longer a tenable position; I shall not repeat their arguments here. Is there an equally clear and compelling argument against at least the strongly reductionist programs which dominate much of the literature in the

neurosciences? Yes. Bracketing for the moment the causal question, I would argue that there is a *difference in kind* between physical explanations of thoughts, feelings and emotions on the one hand, and explanations of those ideas in their own terms on the other. Thoughts have a quality which philosophers (following Husserl) call *intentionality*. The simple definition of intentionality is *aboutness*; it is the characteristic of referring to something else. The referring relationship is intrinsically different from the causal relationship, where A causes B to occur. Causal relationships are clearly physical; they are the bread and butter of the physical sciences, whereas the reference relationship—which we all employ whenever we speak *about* something—works according to a vastly different “logic.” Brian Cantwell Smith states the difference graphically:

Reference—plain, ordinary, vanilla reference, of the sort out of which even the most trivial conversation is made—is manifestly able to leap amazing gaps in space, time and possibility: backwards to the first 10^{-23} seconds of the universe, forward to the death of the solar system, sideways into other possible worlds (such as to a world where Apple responded positively to Bill Gates’ 1985 offer to license the Mac OS). ... This non-effectiveness [of reference] is in direct and exact contrast to physical causality, which is famously ... proscribed from performing any such fancy long-distance or counter-factual maneuvers. ... You can refer to the sun, I take it, *right now*; it doesn’t take 8 minutes for your reference to reach its destination!¹¹

In reference there is no limitation to the speed of light. The particular nature of intentions and conceptual references helps to explain why identity theories (the alleged identity of mental experience and brain states) are inadequate. Imagine that you could (in principle) know exactly what neurological events occur when Michael is asked to define “justice” and makes a verbal response. Still, these events would never be identical to his definition of justice. Consider the analogy with what goes on in your computer’s processor: knowing all the facts about the on and off states of some 16 million registers is not the same as knowing that your computer is currently solving a differential equation.¹²

3.3 Continuing Differences between the More Moderate Positions

It’s still the case that the two more extreme positions garner most of the popular (and media)

11. See (Cantwell Smith unpublished, 8); this paper provides a brief summary of the broader argument in (Cantwell Smith 1996). Michael Arbib argued correctly in a criticism of an earlier draft that leaping gaps in space and time and other “long-distance maneuvers” is not in itself sufficient to show that reference is non-physical. But if one considers the full range of what reference involves, I think it is clear that it’s more closely associated with the logic of the mental than with the logic of physical-causal explanations.

12. It is true, as Theo Meyering has noted, that this is a special case of the more general divergence between structural and functional explanations. My point is that, although the physical or structural facts may determine the emergence of the mental, the mental is something more than its conditions of origination. (Admittedly, the computer analogy may raise further problems of its own because of differences between human and machine intelligence.)

attention in debates about mind. Let us take it as shown, however, that these views—strong reductionism, or “the identity thesis,” and metaphysical dualism—are not tenable. Do we then find a natural middle position emerging? As nice as that would be, it does not seem to be happening.¹³ For one still finds deep divisions between even the more moderate positions. Thinkers such as Jerry Fodor and Hilary Putnam argue that psychology is (more or less) independent of neural considerations; the neurosciences do not play a major constraining role in doing cognitive psychology (Fodor 1981; Fodor 1990; Fodor 1994; Putnam 1975); similar responses are often given by humanistic psychologists. On the other side, Patricia and Paul Churchland, Andy Clark, William Lycan and others argue that psychology and neuroscience must co-evolve: any genuine progress in psychology will give rise to resultant progress in neuroscience (Churchland 1995; Clark 1993; Clark 1997; Lycan 1996). (Note that “co-evolution” is not a mediating position, since it amounts to the denial of autonomous psychology, or folk psychology, in the sense advocated by Fodor and others.)

Still, we’ve now at least arrived at the playing field on which any fruitful debate of the deeper questions in neuroscience and religion (e.g., religious experience) must take place. Also, note that we have now managed to formulate a clear disagreement. The one side argues that functionalist neuroscience can eventually provide as much reliable knowledge of human cognition as humans will ever get, whereas the other side denies this premise, maintaining that there are other types of knowledge of human cognition. What arguments can the two sides develop on this topic?

Let us start with the latter position. In recent years, a more moderate version of dualism—at least more moderate than Cartesian dualism—has been developed in the work of neuroscientists like Sir John Eccles and in several publications by Roger Penrose (e.g., Penrose 1989; Eccles 1989; Eccles 1994). Penrose does believe that there is something like conscious substance, which is ontologically a different sort of thing than physical phenomena. He also maintains that there is “an essential *non*-algorithmic ingredient to (conscious) thought processes” (p. 404). But even so he is not primarily interested in developing a sharply defined dualist metaphysics à la Descartes. Instead he asks, “What *selective advantage* does a consciousness confer on those who actually possess it?” (chapter 10). In my view, Eccles and Penrose are saddled with dualist dilemmas that admit of no easy solution. Yet clearly they represent research programs that are more scientifically respectable than classical Cartesian dualism.

On the other side one finds a more moderate version of the functionalist/neuroscientific position, the “schema theory” as it has been developed by Michael Arbib (Arbib 1992; Arbib et al. 1987). Arbib is not a reductionist in the strong sense of the word. Schemas are “the basic functional unit of action, thought and perception, a unit whose functionality is distributed—in the first instance—across the networks of the individual human brain” (Arbib unpublished, 6). He has also defined them as a “crystallization of some body of experience within a local situation” or simply as

13. Those who have listened in on such debates in the past know that many of the arguments involve members of one moderate and viable research program accusing their opponents of actually holding one of the extreme positions just cited. Clearly, such comments generate more heat than light.

“parallel distributed adaptive computation.” A schema can be “an internal structure or process (whether it is a computer program, a neural network, or a set of information-processing relationships within the head of the animal, robot or human),” or it can be “an external pattern of overt behavior.” These two basic types of schemas can give rise to a “social schema, a schema which is held by the society en masse” (Arbib 1999, 429).

What is interesting about making the schema concept basic for neuroscience is that it is a logical structure which *could* be given either a causal/functionalist or an emergentist interpretation. If one looks at schemas solely in a causal fashion, however, as summaries of causal mechanisms, then eventually they must be reduced down to the basic causal units of neural activity, neuronal firings; and this is precisely what Arbib does. Thus he writes in his recent paper that they are “functional units,” that is, “composable units of brain function/neural activity” (Arbib 1999, 429).

But note that schema theory is a logical device which could in principle be used in a more holistic fashion. One could speak of schemas as phenomena which emerge only at higher levels, when one abstracts from many of the composite parts. For example, cells are schemas—complex wholes—but they are also existing things in their own right. Likewise, my awareness of an orange, or of a situation of injustice, is a highly abstract phenomenon which includes, but goes beyond, countless observations, neural traces, composites, and other influences. Imagine that we gain massive understanding of the workings of your dog’s brain; imagine that our efforts at predicting your canine’s behavior succeed beyond our wildest expectations. Would this prove that your dog does not have subjective experiences (*qualia*) such as fear, concern or affection, or that these qualia do not play any causal role in her actions? Even vastly successful neuroscience thus leaves open the key questions about the mental life. It is *these* questions, not progress in neuroscience per se, that are of life-and-death concern for those interested in the claims of religion.

So I advocate a kinder, gentler schema theory. We need a study of mental phenomena which allows us to focus on higher-order units as (sometimes) genuine existents, not just composites of the parts of which they are composed. In particular, it’s necessary to think persons as distinctive units of activity, as agents capable of forming intentions, making references, and having subjective experiences in the fashion described above. *We therefore need a “science” of the person of which neuroscience is one, but only one, contributing part.* Such a study of the emergent person is genuinely holistic, however, only if it retains a place for speaking of one higher-order event (e.g., a thought or *quale*) causing another without insisting that the whole story can be told in terms of neuronal firings. Arbib, in his well-known work in neural modeling, does not give adequate place to this possibility, though I think schema theory leaves room for it. Still, the crucial fact is that Arbib and others employ a logical framework which could in principle be read *either* in a holistic *or* in an atomistic fashion.

3.4 Closer to a Mediation: Information Biology and Virtual Reality

The last section argued that moderate dualist theories of mind on the one hand, and theories of the

mental as “composable units of brain function/neural activity” on the other, represent positions that are still too far out along the spectrum of positions on the human person. This is not a straw-man dismissal; both are sophisticated positions, and one can see what features in the contemporary study of psychology would drive the authors to their diverse positions. Nonetheless, I believe the strongest theory of the human person lies in between these views. The question then becomes: What view of the self starts neither with theological claims as “obviously given” nor with the “obvious ultimacy” of neuroscientific explanations? What would such mediating position look like?

First, the successful theory will have to grant what we know already from their experience in the world: that our thoughts are not found apart from the functioning of brains, and that damage to the brain can modify or eliminate subjective experience. At the same time, I have argued, the answer must allow for the emergence of mental phenomena and for mental causation. The resulting view will therefore begin the line of causation at the physical level, in a manner similar to the schema theory of Michael Arbib, but at the same time it must insist that a line of causal influence can also be traced (in the appropriate way) *among* the highly complex and abstract “schemas” that we call mental phenomena.¹⁴ Any adequate theory of the human person will have to understand the effect of interactions with the surrounding environment upon mentality, while at the same time doing justice to the irreducible subjectivity of experience.

With regard to the former requirement, the field of information biology has begun to comprehend the way in which all organisms exchange information with the environment around them. In *The Tree of Knowledge: The Biological Roots of Human Understanding*, for example, the biophysicist Maturana and the cyberneticist Varela describe the “structural couplings” that arise between an organism and its surroundings (Maturana and Varela 1992). The organism cannot be decoupled from its environment without dying. The feedback loop that exists between environment and organism is more than just an incidental connection between it and its world. Instead, the way in which those links are set up are physically *and ontologically* constitutive of the organism itself. This is certainly a far cry from a dualist position, where the soul is essentially different from (or uncoupled from) its physical world! At the same time, the information that arises out of these links, and the influence that one’s understanding subsequently exercises via the body on the world, require *a new level* of explanatory concepts. Information-processing agents bring the dimension of subjectivity as one element in this biologically mediated two-way interaction with the environment.

The subjective element in this two-way semantic connection is expressed in some recent experiments in 360° virtual reality. For example, in one well-known artwork in this genre, Char

14. I admit that one can speak of causal influences among ideas, and of ideas on the brain, only in a sense that diverges from the standard use of the term “causality” in science. Here (as in the perplexing “non-locality” results in quantum physic) we need nothing less than a new theory of causality. This theory must supplement the so-called efficient causality on which modern science has been based with a way of speaking of the “causal” influence of form or structure, of function, of information, or of the whole on its parts—yet without falling back into the four-fold causality of medieval metaphysics (formal, final, efficient and material) and the pre- or anti-scientific mindset that it fostered.

Davies' "Osmose," the participant dons a head helmet and a special suit and enters into what seems to be a clearing that surrounds him. He is able to rise in the clearing by breathing in and to lower himself by breathing out; movements are made by gentle leanings from side to side. In the world through which he now "floats" the edges of objects are unclear, and he is able to move through, above, and below the trees and plants at will. This new set of structural couplings between "mind" and "world" can have a profound effect on participants.¹⁵

What occurs when one is in a 360° surround-sound virtual world? It seems clear that the experience gives to subjects a new sense of being embodied—they actually *are* embodied in a different way, thanks to the computer interface. This is why some describe the experience, even months later, as being "within me." New mental experiences *arise out of* one's being given new structural couplings with the world, altering one's mental experience as a result. (Similar mental transformation can arise out of the physical changes associated with drug experimentation, brain disease, or amputation—certainly more brutal forms of cognitive alternation!)

But the lesson to be drawn from these transformations is not the functionalist-reductive one that some neuroscientists would have us accept. For the transformation, though physically dependent, is a *mental* transformation, one whose explanation involves (among other things) psychological concepts. One's particular experience will rely on one's particular set of structural couplings with the world—and, in the case of brain damage, it may be altered by damage to receptors and processing regions—but it is also (irreducibly) about the new mental state that is caused. Equally importantly, these states in turn give rise to a new manner of being embodied in the world and to a new manner of acting causally upon the physical world.

This is a key insight; let me generalize it. The causal line seems to move "up" from the physical inputs and the environment to the mental level, then *along* the line of mental causation—the influence of one thought on another—and then "down" again to influence other physical actions, to make new records and synaptic connections within the brain, to produce new verbal behaviors, and so forth. This view is monist, not dualist: there is only one physical system, and no energy is introduced into that system by some spiritual substance external to it. At the same time, it seems, subsequent states of the entire system cannot be specified without reference to the causal influence exercised by the higher-level phenomena.

In a famous thought experiment by the Harvard philosopher Hilary Putnam, the reader is asked to imagine a "brain in a vat," a brain that has been removed from its body by a team of scientists and kept alive in a vat (Putnam 1981). The imaginary scientists have re-established all of the myriad links that the brain had to its body, replacing them with computer inputs which exactly simulate the body and the environment that the person had experienced before the removal of his brain.

15. Some persons emerge from 15 minutes in this virtual world deeply touched, and (by their reports) sometimes profoundly transformed. Some report that they later experience being embodied in the non-virtual world in a different way, and a few have said, "I am no longer afraid of dying."

The thought experiment may suggest more than Putnam intended, however. Its practical difficulty, verging on inconceivability, underscores how the mental life is highly, even extremely, dependent on our structural couplings with the physical world. The brain possesses an incredibly large number of receptors for information from other parts of the body, and interrelates them with an amazing 10^{14} synaptic connections. Our mental experience is conditioned beyond what we can imagine by the body's vast input to the brain and by the complex way in which the brain processes it. And yet *there is a subjective experience of that world which is different from those physical inputs* and which in turn helps cause the variety of the outputs which constitute our action in the world. (Obviously the brain in the vat would have to be given not only massive inputs, but also the impression that it is acting within the world if it is not to "know" that it has been so rudely imprisoned!) The language of mental impressions, intended references, and mental causes is an irreducible part of the full story, just as in a virtual reality chamber the full story includes not only the new physical inputs to the brain, but also the irreducibly mental dimension of the experience—the transformed mental "place" that arises out of this new virtual-physical surrounding.

4. Emergentist Non-Reductionism

4.1 *Toward a Theory of the Person*

The study of the human person therefore involves not only all the knowledge we can glean about the brain and its workings, but also study of the emergent level of thought, *described and explained not only in terms of its physical inputs and nature, but also in terms intrinsic to itself*. My first task has been to argue for the existence of both levels, and to understand the way in which the mental emerges out of the physical. The second task is to begin to integrate these two levels. What is the best framework for doing this? I suggest beginning with the notion of the human person as *psycho-somatic unity*. Humans are both *body and mind*, and both in an interconnected manner. How does this work?

It is not difficult to describe what is normally connoted by the word "person." A person is one who is able to enter into human social interaction: praising your tennis partner, planning your dinner party for next Friday, carrying out your intention to graduate from college by next May—and being aware of (at least some) other humans as moral agents who have value and rights equal to your own. These are concepts of personhood that are basic to research in the social sciences (psychology, sociology, and cultural anthropology); they are reflected in the literature of various cultures around the world, as well as in multiple religious traditions. Of course, there are many questions that still leave us unsure: when does personhood start? Does it demand a metaphysical basis, such as the introduction of the soul or person-substance? Does it develop and end gradually? Can it be effaced within a human being? Is it a legal or social fiction, or a metaphysical reality? Such broader philosophical questions are crucial to the complete definition of personhood and hence part of the discussion that neuroscientists and theologians must have if they are to find any common ground at

all.

Personhood is therefore a level of analysis that has no complete translation into a state of the body or brain—no matter how complete our neuroscience might be. Of course, it presupposes such states; yet personhood represents an explanatory level that is distinct from explanations at the level of our “hardware.” As Brian Cantwell Smith writes:

First, you and I do not exist in [physical explanations]—*qua people*. We may be material, divine, social, embodied, whatever—but we don’t figure *as people* in any physicist’s equation. What we are—or rather what our lives are, in this picture—is a group of roughly aligned not-terribly-well delineated very slightly wiggling four-dimensional worms or noodles: massively longer temporally than spatially. We care tremendously about these noodles. But physics does not: it does nothing to identify them, either as personal, or as unitary, or as distinct from the boundless number of other worms that could be inscribed on the physical plenum ...(Cantwell Smith unpublished, 3)

The languages of physics and of personhood only partly overlap; one cannot do justice to the one using only the tools of the other. To give a purely physics-based account of the person is like saying that, because a club or church cannot survive without being financially viable (e.g., receiving income from some source), it *just is* the economic unit which economists describe in terms of income and expenditures. The confusion, one might say, is a confusion of necessary and sufficient conditions. A living body and a functioning brain are *necessary* conditions for personhood, yet the wide discrepancy in the “logic” of the vocabularies suggests that they are not *sufficient* conditions. Personhood is not fully translatable into “lower-level” terms; persons experience causal and phenomenological properties (*qualia*) that are uniquely personal.

4.2 Separating the Questions of Science and Ontology

But is this answer permanent or temporary? What if, some time in the future, neuroscience succeeds beyond our wildest imaginings? What if we are some day able to model human behaviors precisely in complex computational machines? Won’t we have shown that personhood is best understood as (something like) a sufficiently complex software system running on the right sort of hardware?

I don’t think so. The debate between physicalist and nonphysicalist views of the person, after all, is not only about science; it is also about what actually or really or finally exists. We must ask: are the properties measured by natural scientists—and recall that we have defined physicalism in terms of the methods of physics—the only sorts of properties that this particular object in the world has? In debating the issue it is important to distinguish the ontology of the phenomena (i.e., of the world as we experience it) from the ontology of the *best explanation* of the phenomena. A cultural anthropologist, for example, might note that the subjects of her study report discussions with the spirits of animals and give explanations of her arrival in their village which conflict with the world as she experiences it (e.g., she is the embodied spirit of one of their ancestors). In *describing* their beliefs, she suspends judgement on their truth, attempting to be as accurate as possible in re-

presenting the world as they see it. In her explanations, however, she will feel free—indeed, it is required of her—to offer explanations which use an ontology (an account of what really exists in the world) that may diverge widely from their own.

The key question under debate, then, is the question of how much of subjective experience or “folk psychology” is irreducible, that is, how much of it actually belongs in a correct explanation of human experience. Some theorists defend an explanatory ontology that consists of brains and other physical organs and their states alone. At the opposite end, others argue that only minds exist, or that both minds and bodies represent primitive substances, defined as radically different sorts of things. Still other thinkers (e.g., social behaviorists) hold that both brains and their social contexts exist, that is, both brains and whatever things we are committed to by an account of social contexts. The view to be defended here, *emergentist supervenience*, holds that brains, social context, and mental properties exist; which means (if I am right) that the correct explanatory ontology has to introduce at least three levels of “really existing properties.” Yet more extensive ontologies are of course available, such as those involving the real existence of ethical predicates, religious predicates, and various religious beings or dimensions. But nothing in emergentist supervenience immediately commits one to other types of properties than the mental.

4.3 Emergentist Supervenience

I agree with several of the other authors in this book that the philosophical notion of supervenience is especially attractive as a bridge framework when discussing neuroscience and the person. Simply put, supervenience grants the dependence of mental phenomena on physical phenomena while at the same time denying the reducibility of the mental to the physical. Note that supervenience is about properties or groups of phenomena, and not about relations between substances (and the ontology that supports them).

Supervenience might be defined as follows:

B-properties *supervene* on A-properties if no two possible situations are identical with respect to their A-properties while differing in their B-properties.¹⁶

The early uses of the concept of supervenience described the way in which ethical judgments are dependent upon certain physical states and yet not reducible to them. The notion made its major entrance into the mind/body debate in the early article “Mental Causation” by Donald Davidson. Davidson writes,

Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respects, or that an object cannot

16. (Chalmers 1996, 33). Contrast this definition with Chalmers’ definition of *logical supervenience*: “B-properties supervene *logically* on A-properties if no two *logically possible* situations are identical with respect to their A-properties but distinct with respect to their B-properties” (Chalmers 1996, 35).

alter in some mental respects without altering in some physical respects. Dependence or supervenience of this kind does not entail reductibility [*sic*] through law or definition: if it did, we could reduce moral properties to descriptive [ones], and this there is good reason to *believe* cannot be done (Davidson 1980, 214).

In the accounts examined so far, there is a direct and full relationship of dependence between the mental and the physical. We might call those views *strong supervenience* in which the physical determines the mental in its emergence and in all its subsequent behavior. Bruntrup writes of the strong supervenience relation, “Micro-properties determine completely the macro-properties (micro-determinism). ... If mental properties are macro-properties in this sense, they are causally inefficacious qua mental properties” (Bruntrup 1998). In this construal of the physical/mental relationship, for example, one might hold that there are general physical laws such that, if they were known, the occurrence of any given mental event could be predicted from a thorough enough knowledge of the brain, its structure, and its past and present inputs.

There is a certain inherent tension in strong supervenience, however. As Jaegwon Kim, one of its best known (former) advocates, admits, “nonreductive materialism is not a stable position. There are pressures of various sorts that push it either in the direction of an outright eliminativism or in the direction of an explicit form of dualism” (Kim 1994; cf. the more recent treatment in Kim 1998). One of the reasons for this instability is that such a position appears to leave no room for genuine mental causes; all the determination of outcomes seems to flow from the bottom (the physical substratum), leaving no “room for play” for the mental actually to do anything. At worst, mental phenomena become mere epiphenomena; their reality is bought at the cost of causal impotence.

So the question becomes: Can any framework that is consistent with what we know today about the brain, and with what we may reasonably be expected to *come* to know, also be consistent with a real causal influence of mental phenomena? Not only folk psychology, the common-sense way of speaking of human persons, depends on a successful theory of mental causation, but the viability of (at least traditional) theological claims does as well. Strong supervenience theories might suggest how religious beliefs and experiences could arise. But however much the *function* of religious beliefs might be incorporated in such accounts, their *truth* could not be. There would be no place for religious insights *as correct* to alter behavior, and definitely no role for any influence of a disembodied divine force on the world. The supervenience concept seemed to offer the sort of framework required for drawing the links between the brain sciences and the mental life that we experience. But strong supervenience conflicts both with folk psychology and with theology.

Is it possible, then, to formulate a “weaker” version of the dependence relationship? Suppose we define *weak supervenience* as the view that, although physical structures and causes may determine the initial emergence of the mental, they do not fully or solely determine the outcome of

the mental life subsequent to its emergence.¹⁷ This view amounts to a dependence of genesis, since it grants that the origins of mentality can be traced to the physical conditions without which there would be no mental phenomena. But it does not grant a full, bottom-up determination of the mental by the physical—hence the “Insufficiency Thesis” defended above—even though the degree of bottom-up influence will certainly far exceed our present knowledge. Weak supervenience thus retains the central tenet of supervenience theory: the mental is dependent on, yet not reducible to, the physical. One reason for choosing weak over strong supervenience is the belief in mental causation: there are genuine mental causes that are not themselves the product of physical causes. The causal history of the mental cannot be told in physical terms, and the outcome of mental events is not determined by phenomena at the physical level alone.

Weak supervenience is the stepping-off point for an emergentist theory of supervenience, and thus an emergentist theory of human personhood. The background for emergentist supervenience comes from the British Emergentists in the 1920s and ’30s. As Jaegwon Kim notes, the early emergentists held “that the supervenient, or emergent, qualities necessarily manifest themselves when, and only when, appropriate conditions obtain at the more basic level; and some emergentists took great pains to emphasize that the phenomenon of emergence is consistent with determinism. But in spite of that, the emergents are not reducible, *or reductively explainable*, in terms of their ‘basal’ conditions” (Kim 1993, 138). Lloyd Morgan thus appeared to use “supervenient” as an occasional stylistic variant of “emergent” (so Kim 1993, 134).

In a recent article, O’Connor has defined property emergence in a more careful manner:

Property P is an emergent property of a (mereologically-complex) object O iff:

- (1) P supervenes on properties of the parts of O;
- (2) P is not had by any of the object’s parts;
- (3) P is distinct from any structural property of O.

But after those three conditions we come to the big break in the philosophy of mind, the question that Kim calls “arguably the central issue in the metaphysics of mind”: the question of mental causation (Kim 1993, xv). O’Connor formulates it this way in his final premise:

- (4) P has direct (“downward”) determinative influence on the pattern of behavior involving O’s parts (O’Connor 1994, 97f).

It is not difficult to provide a formal definition of emergence in this sense: “F is an emergent property of S iff (a) there is a law to the effect that all systems with this micro-structure have F; but (b) F cannot, even in theory, be deduced from the most complete knowledge of the basic properties of the components C_1, \dots, C_n ” of the system (Beckermann 1992, 104). Note that emergent properties of this sort are genuinely novel. As Bruntrup writes, “Even if all the physical facts have been fixed, the emergence of consciousness is not implied with nomological necessity. ... The

0. Note that strong and weak supervenience has been used in other (not always consistent) ways in the literature. I choose to run the risk of terminological confusion because of the particular appropriateness of these two terms to the position defended in the text.

existence of emergent properties could not be predicted by even a perfect knowledge of the underlying physical facts alone” (Bruntrup 1998, 104).¹⁸

A property is thus emergent only if laws cannot be formulated at the lower level that predict its occurrence and behavior, say as a boundary condition of other well-established laws at that level. If for example we can relate the levels with the same bottom-up precision with which we can formulate the necessary physical conditions for the existence of conductivity or elasticity, then we do not have emergentist supervenience. A set of phenomena is designated as emergentist only when an exhaustive description of the underlying physical state of affairs, although necessary, is not sufficient for explaining the emergent properties. Thus an emergent condition seems to be implied in Leslie Brothers’ explanation of human social behavior in terms of “the representation of the generalized other” and the irreducible nature of first-person language—assuming that she means these terms to refer to a genuinely psychological reality that is something more than, and not just a different manifestation of, the underlying physical realities. One would also need to use the language of emergence if qualia (human subjective experiences, such as seeing red or being in love) are, at least in part, self-explaining.

I believe that emergentist supervenience offers the philosophically most adequate framework for conceptualizing mental properties in human persons. Does emergentist supervenience also offer a view of the person that is more compatible with theology than strong supervenience as defined above? If true, would it represent, from the standpoint of theology, a better bridge principle? Clearly the answer is yes. Presumably theologians would have many more things to say about emergent properties and their source and ultimate purpose. They might also attempt to offer theologically based explanations of why the biological world could or would give rise to such emergent properties. Two caveats, however: when speaking in this way, theologians do not speak as scientists, and the status of such language vis-à-vis any presently available empirical verification should be made fully clear. Also, there is nothing in emergentist supervenience that *requires* a theological interpretation; it is not a form of natural theology. Emergentism is, in my view, a necessary condition for a theological interpretation of the human person, but it is emphatically not a sufficient condition for a theological anthropology.

Coming from the viewpoint of science, one might worry that such a position closes off research and hence progress in neuroscience. Does it introduce a constraint on the work of empirical scientists? I would argue not. Emergentists may have an equally vivid interest in knowing more about actual brain functions and in seeing neural explanations extended as far as possible. It is just that they wager that the “as far as possible” does not extend as far as an exhaustive explanation of the mental—unless part of that explanation is given in irreducibly mental terms! Talk about the subjective experience of being in love or the sense of self-awareness is irreducibly mental; such

18. Note that in this article Bruntrup does not accept the position that I am defending.

phenomena exercise a type of causal influence of their own.¹⁹

5. Persons and Explanatory Levels

By exploring the family of positions that eschew both dualism and strong reductionism, this paper has focused on that range of positions that seek to do justice both to neuroscience and to the human experience of personhood. Following contemporary usage, I have characterized these as the family of supervenience theories. We discovered that the same tension arises within this family as was present in the old *dualism versus reductionism* debate. One either does or does not accept the Sufficiency Thesis, the view that the causal explanations of human behavior will ultimately be given in neuroscientific terms.

Since both sides of this new debate accept the supervenience label, one might suppose that the ambiguity lies in the term itself. And indeed this is exactly right: one finds in the literature at least three different ways of characterizing the relation of mental to physical. All three presuppose that mental phenomena represent levels of complexification that depend on lower, more simple levels, yet that are in some sense not fully reducible to those lower levels:

(1) The more complex level could be related to the lower level by a clear set of laws (call it *nomological supervenience*). In the present volume this appears to be the position of Theo Meyering, for whom the paradigm supervenience relationship is expressed by phenomena such as elasticity and conductivity. These are phenomena that are well understood scientifically in terms of the behavior of the particles making up the physical system in question, although the supervenient properties cannot be fully expressed except at the level of the set of particles as a whole. Nomological supervenience is also visible in the work of R. M. Hare, who says explicitly that “supervenience brings with it the claim that there is some ‘law’ which binds what supervenes to what it supervenes upon.” For Hare such laws are necessary conditions for supervenience: “what supervenience requires is that what supervenes is seen as an instance of some universal proposition linking it with what it supervenes upon” (Hare 1984, 3).

(2) The higher level could have all of the attributes listed in (1), yet without the condition just expressed by Hare, which we might call the “nomological condition.” This second position is best known in the guise of what Donald Davidson calls “anomalous monism.” Davidson holds that “mental entities (particular time-space and space-bound objects and events) are physical entities, but ... mental concepts are not reducible by definition or natural law to physical concepts” (Davidson 1995, 3; cf. Davidson 1980). Davidson disputes the lawlikeness of mental events: mental events are of a different type than physical events, although there is a token identity of every mental event with a physical event. Still, in other respects his view stands fairly close to (1). Certainly he does not speak of mental phenomena as genuinely emergent. He insists only that at least one portion of the

19. Indeed, wouldn't it be a strange thing for a neuroscientist to find herself in the position of denying with passionate subjective conviction that there is any such thing as a *force* of subjective conviction?

physical world does not admit of the kinds of causal explanation by means of natural laws that science has been successful in formulating in so many other areas. Yet no emergentist conclusions should be derived from this particular failure of law-like explanation, Davidson seems to say; the mental simply obeys different constraints than physical laws, such as the unique constraint of rationality.

(3) The final type of supervenience is the one that I have been defending. It finds in mental phenomena and their dependence on the physical a supervenient relationship not unlike that accepted by the other positions. Yet it also finds grounds in the nature of this relationship to support the ontological hypothesis that mentality represents an emergent level. That is, without questioning the dependence on the physical, it understands mental properties to be different in kind from the properties that one observes at lower levels and to exercise a type of causal influence unique to this new emergent level.

5.1 Minimalist Emergence

It is important to note that a majority of philosophers writing in the field still advocate either (1) or (2). I believe that this reveals a shared sense of what is at stake in the present debate. If one wishes to avoid talk of self-consciousness (say, in the causal sense used by the German Idealists), or God-talk, or an opening for any other such religiously-tinged predicates, then one must insist that the mental be understood fully in terms of the physical world. By contrast, if one finds in the mental some sign of a new type of phenomenon within the world, then one has thereby introduced at least the *possibility* that there is something inherently mental or spiritual within the one world that we find around us. Clearly this possibility would represent an opening to theology that is of great significance to both sides. If one wishes to avoid such openings, then one must be sure at every cost that the mental is not interpreted in an emergentist sense. Conversely, it seems that those with theological interests—and with some motivation to integrate these interests with their understanding of science—will need to develop a theory of humans and their mental life that is either emergentist or establishes the same sort of minimal opening that emergentism defends. These are the stakes that make the present discussion and the present book of such overwhelming importance. It is perhaps not too much to say that this debate about the human person expresses the crux of the battle between physicalist naturalism and its opponents today.

It is also a debate with no easy resolution, as we have seen. “Opening” means possibility, not proof; no one is talking of conclusive demonstrations here. One could easily accept emergentist supervenience and deny the truth of theism or religious belief in all its forms.²⁰ Still, even in this incredibly circumscribed form, emergentist supervenience presents one with what philosophers call a

20. Indeed, as I show in the final chapter of (Clayton 1998a), emergentist supervenience stands in a certain tension with traditional theological belief, which asserts a dependence of the physical on the spiritual. Either the dependence of the mental on the physical that I have defended must be corrected from another source, or it will require significant revisions in traditional Christian belief.

“forced choice.” If one holds that all mental phenomena are only expressions of physical causes or are themselves, at root, physical events, then one has (at least tacitly) advanced a theory of the human person that is pervasively physical. It then becomes extremely unclear (to put it gently) why, *from the perspective of one’s own theory of the human person*, a God would have to be introduced at all (except perhaps as a useful fiction). If a theologian espouses physicalism, she may be forging an alliance with the majority worldview within the neurosciences, but she may also be giving up the most interesting rapprochement between theology and the sciences of the person just as she approaches that debate’s most decisive issue. By contrast, to introduce a soul-substance at this crucial juncture would be to abandon the debate altogether, for that move, almost by definition, leaves no common ground with natural science. Here I have argued not that supernatural souls exist but rather that human action reflects a type of mental causation that is something more than physical. This claim, minimalist as it is, may just be the necessary condition for a theology that is anything more than metaphorical. Theologians stand before their Rubicon and must either cross or not cross.

My strategy has been to map out a crossing where the river is most narrow (why add any unnecessary distance when the crossing itself is already difficult enough?). This helps explain why I have broken with dualist thinking and moved as far as possible in the direction of the natural sciences by arguing that:

- * mental predicates represent a type of property, not a new form of substance;
- * mental causation does not involve the addition of new energy into physical systems;
- * mental processing does not occur without concurrent physical activity. Indeed, changes in brain structure and function (brain disease, lesions) have important and predictable effects on mental functioning;

- * one’s overall ontology should be monist. There is only one natural order, although it includes many different types of things. Mental causation is not supernatural; it is natural. It is thus amenable to explanation in this-worldly terms, although at least part of the explanation will need to employ irreducibly psychological concepts. I have not pleaded for supernatural interventions, nor have I construed mental functioning in any way analogous to the classic supernaturalist notion of intervention from outside. To put it bluntly: though there may be divine action on analogy with the action of embodied persons with the world, I have left no place for miracles in the sense of a countervailing of natural law.²¹

I imagine, one final time, the objection, “Well, you have wagered against neuroscience, have you not?” The critic might object that I have introduced, if not a “God of the gaps,” then at least a “mental causation of the gaps.” Isn’t the more scientific response to expect that law-like explanations will eventually be possible “all the way down”—until all phenomena in the natural world have been explained from the bottom up? Doesn’t “wagering” in this way amount to betting against science, and thus blocking the road for scientists? Indeed, isn’t the success of science heretofore good reason

21. The details of what divine action would look like in this context are spelled out in (Clayton 1998a).

to conclude that bets on my side are backward-looking, obscurantist, and in general inhibitors to further scientific progress?

No. These well-worn objections tell against dualist positions, but they beg the question at dispute between supervenience theorists. The reason it is absurd to postulate occult forces in the physical world (or “vitalist” forces in the biological world) is that we have learned that these realms operate in a fully law-like manner *based on explanatory successes in the relevant sciences*. What is really at stake in the present book is the question whether human persons are analogous—whether they can be exhaustively predicted and explained in a “bottom-up” manner. I have argued that we have good evidence to think not. Indeed, the hierarchy of the sciences itself offers evidence of principles which are increasingly divergent from “bottom-up” physicalist explanation.²² Functionalist explanations play a role in the biological sciences (from cell structures through neural systems to ecosystem studies) that is different from the structure of explanation in fundamental physics, just as intentional explanations play a role in explaining human behavior that is without analogy at lower levels.²³ These emerging orders of explanation may also involve an increasing role for top-down explanations. Thus, for example, DNA embodies in its very structure the top-down action of the environment on the molecular biology of the human body. In intentional explanations it is even more clear that the goal for which the agent acts, or the broader context within which she understands her actions, influences the particular behaviors or thoughts. An emergentist view of the person is thus not an argument against science but rather consistent with the pattern that we find emerging in the natural hierarchy of the sciences

5.2 Emergence, AI, and the Social Sciences

This last point is important enough to bear restating: the case for emergent mental causation is not by itself a case for the existence of God, divine action, an eternal soul, or life after death; it is not directly a theological conclusion at all. Indeed, in some ways it might seem to be an *anti*-theological conclusion, because it understands mental phenomena to be “of a piece” with physical phenomena, and because the supervenience relationship asserts a dependence of the mental life on its physical basis—indeed, a high correlation between physical causes and mental effects—which is on the surface inconsistent with many parts of Christian teaching. To accuse this view of being a cheap theological concordism is to neglect all of these sharp differences. I have argued only that human mentality is an emergent feature of a very complex biological structure, the human brain, in its

22. I cannot review the entire argument here. It is powerfully laid out in (Peacocke 1993).

23. These emerging orders of explanation may also involve an increasing role for top-down explanations. Thus, as George Ellis has pointed out (in private communication, and more recently in a February 2000 contribution to the META listserve), DNA embodies in its very structure the top-down action of the environment on the molecular biology of the human body. In intentional explanations it is even more clear that the goal for which the agent acts, or the broader context within which she understands her actions, influences the particular behaviors or thoughts.

interaction with its environment.²⁴ This conclusion is not a No to science and a Yes to faith. Rather, it suggests that one will have to supplement the neurosciences with another set of sciences, say the human sciences, before one can provide the full explanation of that particular part of the natural world which is us.

Consider the parallels with the Artificial Intelligence (AI) program in computational theory and computer science. The challenge of the Turing test was to build a computational device whose outputs could not be distinguished by human agents from human outputs. In seeking to meet the test, computer scientists first worked at what is now called (by its critics) “brute force” AI. The hope was to achieve a functional similarity to human outputs by means of sheer computational power. However, already by the time of Deep Blue, the chess program that beat Kasparov, a variety of other techniques had been introduced, effectively supplementing brute-force computation by a combination of heuristics and higher-level criteria. In the process of leaving behind brute-force AI, the very understanding of computational theory has been transformed. It has now been stretched— rightly, in my view—to include fundamental questions in semantics and the theory of meaning, so that theories of computation can now include holistic considerations such as the impact of broader semantic systems, contexts and applications that go well beyond the actual computations in question. In very recent years, in what appears as a natural next step, some thinkers have even migrated from computational theory to fundamental ontology, the debate over how worlds are first constructed by means of information (cf. e.g. the detailed argument in Cantwell Smith 1996).

One notes a certain irony here. The battle began over what is the most scientific approach to take in the study of humans. Those who offer a purely physicalist account of human mental predicates claim for themselves the laurel of scientificity and accuse their opponents of obscurantism. Yet according to the opposing position it will actually be more scientific to *deny* that human intentional actions can be explained as law-like phenomena—*if*, as emergentists and anomalous monists believe, the phenomena in question are actually more than physical in nature.

The ongoing debate about the nature and methodology of the social sciences recapitulates (and sheds some helpful new light on) the discussion to this point. The two opposing camps appeal to the two warring fathers of modern social science, August Comte and Wilhelm Dilthey. Comteans argue for a predominantly natural scientific approach to the social sciences, allowing no in-principle gap between them and the natural scientific study of the human organism (Comte 1988). Present-day Diltheyans maintain that the object of study to which the human sciences are devoted is significantly different from the natural world. The natural world can be grasped using *causal* patterns of explanation, because such events really are the product of a series of causes. But human actions require the method of *Verstehen* or *empathetic understanding*, for human beings are subjects who are engaged in the project of making sense of their own world. Intentional actions can be understood

24. Leslie Brothers emphasizes this dimension of sociality in (Brothers 1997). On her account, the mental world grows (for instance) out of the sort of brain-brain interaction that we call conversation.

only in terms of the logic of intentionality: wishing, judging, believing, hoping.²⁵

The battle continues. A new round was launched by the successes of behaviorist social science, by Abel's oft-cited Comtean manifesto for positivism in the social sciences²⁶, and more recently by the rapid advance of the neurosciences; shots were then returned by humanist psychologists and by more hermeneutically inclined theorists (see esp. Gadamer 1975). At the same time, analytic thinkers have carefully stressed the difference between explanations of human intentional actions and causal explanations of occurrences in the world, as in Georg Henrik von Wright's detailed defense of the logic of intentional explanations (von Wright 1971). Whereas Carl Hempel tried to subsume the explanation of human actions under his general model of deductive-nomological explanation, other leading philosophers of science such as Ernest Nagel underscored the unique nature of explanations of social action (Hempel 1965; Nagel 1961). The net result is a clearer sense of what it is that sets person-based explanations of individual and social action apart from causally based explanations.²⁷

6. Separating Theology and the Theory of the Person

What relevance do such general reflections on the philosophy of social science have for the debate between the neurosciences and theology? The latter discussion is often set up as a special case of the more general debate over the independence of science. Anyone who suggests that there might be a limit to the scope of neurologically based accounts of human thought is understood to be doing something to scientific inquiry analogous to what supernaturalists do to physics when they insist that God can break into the natural order at any time, bringing about any result He pleases with no attention to natural causal influences or the requirements of the energy conservation principle.

Yet a moment's reflection shows that the two instances are precisely *not* analogous. The appeal to divine inbreaking and miracles is the appeal to actions carried out by an agent whose existence is contested and who is by definition unique, unlike any other agent. By contrast, explanations in terms of human intentions appeal to experiences that are (presumably) shared by every agent who can read and comprehend the words on this page. Cultural anthropology studies intentions shared by large groups of actors, as does sociology; and even psychology, with its focus on what is uniquely individual, still speaks in terms of shared personality types, motivations, complexes, structures, and pathologies. In all these cases, the explanatory (scientific) goal is to

25. See the excellent summary of Dilthey's thought in (Makkreel and Rodi 1996) and (Makkreel and Rodi 1989). Dilthey used this argument as the basis for his broader theory of the social sciences. The debate was repeated in the work of Windelband and others; see for example (Windelband 1912).

26. See (Abel 1970) on the explanation versus understanding debate.

27. (Giddens 1976) explains the difference in terms of the "double hermeneutics" that characterizes social explanations.

understand and explain the behavior of a large class of agents called human subjects—a type of natural entity with which each of us is deeply familiar, in part through direct introspective awareness. The question at issue, then, is not (in the first place) a theological question at all, but a basic question about human agency: should we expect in the long run that neuroscientific explanations will be adequate to explain human behavior, or do social scientific explanations pick out a type of action in the world which demands an explanatory level of its own? One’s motivation need not be in any way theological in order to defend an account of the human self that includes self-conscious intentions as basic building blocks of human behavior.

This distinction, albeit crucial, is more difficult than it may at first appear. A neuroscientist such as Michael Arbib, for example, might well insist that he too preserves a place for social scientific explanations. He might well (indeed, does!) insist that schema theory as discussed above supplements base-level neurological explanations of human behavior with a higher level of analysis, one which includes those schemas that make up the social world. The difference remains, however. Arbibian schemas are constructs composed out of neurological events and physical events in the world, which are their real foundation. Individual actors may *believe* that things such as societies and their institutions—and ideas such as freedom and responsibility, not to mention divine beings—really exist. But they are mistaken. Eventually, when we understand the physical nature of the human being well enough, we will be able to give a complete account of how ideas such as these came to be constructed. At that time we will leave behind the fictions of their independent reality and return to a purely physicalist account of ourselves and our computational products.

It is a very different thing to argue, as I have, that the social sciences pick out phenomena which, even though they have emerged from the physical world, are causally irreducible. This is an ontological claim, but it is also a claim about what the study of human persons entails, a claim about the form that (at least some) explanations of persons must take. It’s the question that is ultimately at stake in the debate to which this book is devoted.

Interestingly, we have found that the debate about neuroscience and personhood has a sort of fractal structure. When we left behind dualism and strong reductionism, it looked like we would find common ground; but the ground quickly fissured into (e.g.) schema theory and agent-centered theories. When we introduced supervenience theory as common ground between those views, we discovered multiple meanings of supervenience, which mirrored the tensions already encountered at the first two levels. Presumably, if there were space to explore the concept of emergence in greater detail, we would find the same disagreements occurring again at this level. And yet the iterations have helped to give sharp profile to the recurring dispute: are neurological explanations finally sufficient or insufficient?

7. Emergentist Monism

One might ask, “What does it all mean? What kind of ontological position do these emergent properties entail? Is it monism or property dualism or panpsychism? And where does this all leave

theology?”

The ontological view that I defend might be called *emergentist monism*.²⁸ Monism asserts that only one kind of thing exists. There are not two substances in the world with essentially different natures, such as the *res cogitans* and *res extensa* (thinking and extended substance) propounded by Descartes and the Cartesians. But unlike dual-aspect monism, which argues that the mental and the physical are two different ways to characterize the one “stuff,” emergentist monism conceives the relationship between them as temporal and hierarchical.

In one sense, monism is a necessary assumption for those who wish to do science. For instance, we can (and must) assume that the total physical energy of the universe as a whole is conserved. No action that you perform, no thought that you think, can add to the total energy of the system without invalidating calculations based on physical laws. Incidentally, this is the problem with dualism, and with direct interventions into the world by a God who breaks natural laws: if a spiritual cause gives rise to a physical effect, it has brought about physical change without a physical cause or the expenditure of physical energy, and this fractures the natural order in a way that would make science impossible. There could be no scientific study of a world where cups spontaneously fly across the room and objects released from your hand could go either up or down according to spiritual forces. Science does not need full determinism (see the next paragraph). But it does need the world to reflect at least patterns of probability over time.

(Note that monism is not only in the interests of science; one can *also* give theological arguments in defense of monism. Monism makes the assertion that the world is one, that it constitutes a distinct order. Theologians speak of the universe as a whole as *finite*, in order to specify its single ontological status and to contrast it with a Creator whose nature is essentially infinite. Herein lies the theological importance of the phrase, “the unity of nature”: in comparison to the Creator, all things in the universe share a common nature. Theologians have also argued that creatures can only exercise freedom within an ordered world that has an integrity and lawlike structure of its own.)

I don’t care if you want to think of this monism as a *sort* of materialism, but only if you mean by that that the “things” in the world—rocks and computers and persons—are all made out of *some material or other*. What’s crucial is that you develop theories which do justice to the specific qualities that we actually find associated with the various “things” in the world. For example, after Newton we thought that physics presupposed at least the possibility of a fully determinate, and determined, account of the world. But when we found out that microphysical or quantum events simply don’t work this way, we developed an essentially stochastic or probability-based science to deal with them. Likewise, when scientists began to research chaotic “systems,” or systems far from thermodynamic equilibrium, they discovered that they were *essentially* unpredictable (for finite agents). But science did not end; instead, a fascinating new science of chaotic systems has been

28. The term was developed in ongoing conversations with Arthur Peacocke; see (Peacocke 1999).

developed. An equally complex story would have to be told about the convertibility of matter and energy.

Now we come to a *very* complex object in the world: humans. With 10^{14} neural connections, the brain is the most complex interconnected system we are aware of in the universe. This object has some *very* strange properties that we call “mental” properties—properties such as being afraid of a stock market crash, or wishing for universal peace, or believing in divine revelation. On the one hand, to suppose that these features will be fully understood in terms of physics as we now know it is precisely that: a supposition, an assumption, a wager on a future outcome. A deep commitment to the study and understanding of the natural world (which I share with most, and probably all, the contributors to this volume) does not necessitate taking a physicalist approach to the human person—if by that one means that the actions of persons must be explained through a series of explanatory sciences reaching down (finally) to physics, or, more simply, that all causes are ultimately physical causes. (Note that under this definition there could be both reductionist and nonreductionist versions of physicalism.) On the other hand, *for both scientific and theological reasons*, I do not therefore advocate introducing an occult entity, such as Descartes’ soul substance, in order to explain the person. To say that the human person is a *psycho-somatic unity* is to resist both positions. It is instead to say that the person is a complexly patterned entity within the world, one with diverse sets of naturally occurring properties, each of which needs to be understood *by a science appropriate to its own level of complexity*. We need multiple layers of explanatory accounts *because* the human person is a physical, biological, psychological and (I believe also) spiritual reality, and because these aspects of its reality, though interdependent, are not mutually reducible. Call the existence of these multiple layers *ontological pluralism*, and call the need for multiple layers of explanation *explanatory pluralism*, and my thesis becomes clear: ontological pluralism begets explanatory pluralism. (Or, to put it differently: the best explanation for explanatory pluralism is ontological pluralism.)

In her essay in (Russell et al. 1999), Nancey Murphy draws on the work of Ian Barbour and Arthur Peacocke in chronicling the multiple meanings of “reductionism.” Given her definitions, note that an emergentist position rejects causal reductionism, since it accepts mental causes. It therefore rejects explanatory (theoretical, epistemological) reductionism, insofar as mental properties need to be explained using a theoretical structure appropriate to them. At first blush, emergentist monism may *seem* like a version of ontological or metaphysical reductionism, since it breaks with dualism and refuses to postulate non-physical entities such as souls. But emergentists must finally declare themselves opposed to reductionism even with respect to ontological (metaphysical) questions. For their central assertion is that the history of the universe is one of development and process. The one order exists at each stage in its history, but *what it is* that exists is not identical through time. Genuinely new properties emerge which are not reducible to what came before, although they are continuous with it.

What *emerges* in the human case is a particular psycho-somatic unity, an organism that can

do things both mentally and physically. Although mental functions supervene upon a physiological basis, the two sets of attributes are interconnected and exhibit causal influences in both directions. We therefore need a science or mode of study that begins (as a science should) with a theoretical structure adequate to this level of complexity. To defend an emergentist account of the self is not to turn science into metaphysics. Instead, it is to acknowledge that the one natural world is vastly more complicated and more subtle than physicalism can ever grasp. You can *wager* that the *real* things that exist in the world are physical processes within organisms, and that everything else—intentions, free will, ideas like justice or the divine—are “constructs,” complicated manifestations of neural processes. But I’m wagering on the other side. I wager that no level of explanation short of irreducibly psychological explanations will finally do an adequate job of accounting for the human person. And this means, I’ve argued, the real existence and causal efficacy of the conscious or mental dimension of human personhood.

8. Some Potential Objections

I conclude with some objections that might be (or have been) raised against this position.²⁹

8.1 Why not use the label “physicalism”?

As long as one is concerned with the physical sciences and the study of objects in the physical world, why not use the label “physicalism” as the overarching position? Perhaps (my critic might grant) it’s important that scientists countenance causal influences at various levels, and thus also explanations at each of these levels as well. Put differently, perhaps it’s necessary to resist causal and explanatory reductionism. But why resist ontological reductionism? Indeed, what’s the point in arguing about ontology anyway?

Presumably it’s not necessary for working scientists to take a strong position on ontological questions like “what really exists” (though this fact doesn’t prevent some from being vociferous advocates of ontological positions like physicalism and materialism). Moreover, one can work in some areas of the philosophy of science without raising ontological questions. But surely *theologians* hold some important ontological commitments, for they are committed to the existence of a spiritual being or dimension which, while it may include the world, transcends it as well. The biblical doctrine of the image of God (*imago dei*) suggests that something of the spiritual nature of this God is reflected in the nature of human beings (and perhaps in other parts of the natural world as well). Surely this fact commits theologians to an ontological thesis: the thesis that human persons, correctly and fully understood, include a spiritual dimension which, whatever else it is, is more than physical. It is for this reason that theologians, at least, cannot eschew the ontological question and cannot, at the end of the day, be satisfied with the label “physicalist.”

29. Although I have not cited persons by name in what follows, I am again grateful to the various members of the CTNS/Vatican Observatory working group for raising these (or related) criticisms over the two years of the project.

8.2 Is this theory crypto-dualist?

Clearly the position defended here is not a version of substance dualism; there has been no suggestion of mental substances intervening in the physical order. But is it a variant of *property* dualism, the view that, even if there is only one kind of substance, it has two fundamentally different kinds of properties?

Such a criticism rests on a misunderstanding. I have not portrayed a world divided into two distinct types of qualities, but rather a world with a vast array of different types of properties. Though there is no justification for the “dualism” label, the theory could fairly be called *property pluralism*, since it countenances a wide range of properties depending on their position in the complexity hierarchy. In a similar vein, Roger Sperry writes of his own position, “Because it is neither traditionally dualistic nor physicalistic, the new mentalist paradigm [in the study of consciousness] is taken to represent a distinct third philosophical position. It is emergentist, functionalist, interactionist [in the sense that it sees mental phenomena ‘as primarily supervening on rather than intervening in the physiological processes’], and monistic” (Sperry 1983, 165). Once one has grasped the hierarchical structure of the physical world, one can leave the old opposition between physicalism and dualism behind.

8.3 Why not think that full neuroscientific explanations of consciousness will eventually be available? Doesn't the view taken here block progress in neuroscience?

“One should be cautious about wagering against scientific progress!” this critic complains. “Who would have imagined what we would learn through the new scanning technologies, or micro-surgery techniques, or computer modeling? For that matter, who could have imagined 50 years ago what computers would be capable of today, and who can guess what they’ll be doing 50 years from now? Never bet against science!”

But my wager against the Sufficiency Thesis does not stem from underestimating the likely advances in neuroscientific theory or from dreading the coming advances in this field; far from it. Instead, it stems from *limitations in principle* which neuroscientists, philosophers and even some theologians have neglected. Consider a parallel case: when a quantum physicist tells you that she has reason to think that even major advances in her field will not overcome physicists’ inability to know both the location and the momentum of a subatomic particle with full precision, she is not being a pessimistic reactionary, for there are compelling scientific reasons to think this limitation on our knowledge is intrinsic to quantum phenomena. Likewise, there are strong reasons to think that qualia are *intrinsically* subjective experiences. As the immunologist Gerald Edelman (who is certainly no dualist!) writes,

We cannot construct a phenomenal psychology that can be shared in the same way as a physics can be shared. ... What is directly experienced as qualia cannot be fully shared by another individual as an observer. [And later:] There is something peculiar about consciousness as a subject of science, for consciousness itself is the individual, personal

process each of us must possess in working order to proceed with *any* scientific explanation (Edelman 1992, 114, 138).

As in the quantum case, there is something in the case of qualia—an essentially first-person aspect—that makes them irreducible to the third-person scientific perspective. This aspect, which philosophers knew (and all human subjects know?) as consciousness or self-awareness, represents perhaps the single strongest argument on behalf of mental qualities as genuinely emergent in the sense defended in this paper. If Edelman is right, qualia cannot be exhaustively explained by neuroscience because they are the precondition for there being any scientific explanations in the first place.

8.4 Can you have emergence without falling into vitalism, idealism or other scientific heresies?

Analogous to Arthur Peacocke’s paper elsewhere in this volume, I have sought to make the case that emergentist monism is not a quirky or anti-scientific metaphysical position. The strength of (Peacocke 1993) is that it introduces emergent properties not just at the spiritual or mental level, or at the origin of life, but as a pervasive principle running through the hierarchy of the sciences. Emergent phenomena might be seen to occur even at the level of physical chemistry; as the chemist Joseph Earley has recently written, “Chemical combination generates properties and relations that are not simply related to the properties and relations of the components” (Earley 1998, 3). (Of course, one would have to appeal to a broader theory of emergence of the sort defended here to show that chemical phenomena are not merely physical phenomena under a different description.)

What is especially intriguing about the emergentist position is that it makes mental phenomena not an *exception* to the patterns in other sciences but rather yet another instance of them—albeit “higher,” more complex and in some respects stranger than any other properties of the natural world known to us. The vitalists, neo-idealists and (to a lesser extent) the British Emergentists of the 1920s were all committed to a strongly metaphysical position which they brought *to* the biological sciences. This is why critics are justified in dismissing especially the first two positions as in tension with scientific results and methods. By contrast, we are advocating a theory of emergent properties no stronger than is required to interrelate results up and down the hierarchy of the sciences. We claim that this theory does a *better* job of interpreting the connections (and discontinuities) between various scientific disciplines than do any of its competitors.

Of course, one can speculate further about emergent properties at still higher levels (spiritual properties, say), or about orders of reality beyond the natural order as a whole; and surely systematic and philosophical theologians will find it necessary to pursue some of these lines of reflection, as I have done elsewhere.³⁰ But to defend an emergentist theory of mental properties in dialogue with the neurosciences, as has been done in these pages, does not immediately commit one to a full-bodied

30. For example, I’ve suggested some possible threads to pursue in the final chapter of (Clayton 1998a), esp. pp. 257ff. See also (Clayton 1998b), as well as four critiques of my view and a response in (Clayton 1999).

(or: a fully *disembodied*) theology or theory of the supernatural. In one sense I have sought nothing more here than to resist an unnecessarily reductionist interpretation of recent neuroscientific results that would bring them into conflict with those other disciplines (and experiences) that must also play a role, finally, in a full theory of the human person.

References

- Abel, Theodore F. 1970. *The Foundation of Sociological Theory*. New York: Random House.
- Arbib, Michael. 1992. "Schema Theory." In *The Encyclopedia of Artificial Intelligence*, ed. S. Shapiro. New York: Wiley. 1427-43.
- . 1999. "Crusoe's Brain: Of Solitude and Society." In *Neuroscience and the Person: Scientific Perspectives on Divine Action*, ed. Robert Russell, Nancey Murphy, Theo Meyering, and Michael Arbib. Vatican City State: Vatican Observatory Publications. 419-448.
- . Unpublished. "Computing the Self and the Horrors of Humanity."
- Arbib, Michael; E. Jeffrey Conklin; and Jane C. Hill. 1987. *From Schema Theory to Language*. New York: Oxford University Press.
- Armstrong, D. M. 1993. *A Materialist Theory of the Mind*. New York: Routledge.
- Aylward, E. H.; N.B. Anderson; F.W. Bylsma; M.V. Wagster; P.E. Barta; M. Sherr; J. Feeney; A. David; A. Rosenblatt; G.D. Pearlson; and C.A. Ross. 1998. "Frontal Lobe Volume in Patients with Huntington's Disease." *Neurology* 50 (January): 252-58
- Beardsley, Tim. 1997. "The Machinery of Thought." *Scientific American* (August): 78-83.
- Beckermann, Ansgar. 1992. "Supervenience, Emergence and Reduction." In *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, ed. A. Beckermann, Hans Flohr, and Jaegwon Kim. New York: W. de Gruyter. 94-118.
- Brothers, Leslie. 1997. *Friday's Footprint: How Society Shapes the Human Mind*. New York: Oxford University Press.

- Bruntrup, Godehard. 1998. "The Causal Efficacy of Emergent Mental Properties." *Erkenntnis* 48: 133-45.
- Chalmers, David John. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Churchland, Patricia S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge: MIT Press.
- Churchland, Patricia S.; and Terrence J. Sejnowski. 1992. *The Computational Brain*. Cambridge: MIT Press.
- Churchland, Paul. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge: MIT Press.
- Clark, Andy. 1993. *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge: MIT Press.
- . 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge: MIT Press.
- Clayton, Philip. 1989. *Explanation from Physics to Theology: An Essay in Rationality and Religion*. New Haven and London: Yale University Press.
- . 1997. "Inference to the Best Explanation," *Zygon: Journal of Religion and Science* 32 (September): 377-91.
- . 1998a. *God and Contemporary Science*. Edinburgh and Grand Rapids: Edinburgh University Press and Eerdmans.
- . 1998b. "The Case for Christian Panentheism." *Dialog* 37 (Summer): 201-208.
- . 1999. "The Pantheistic Turn In Christian Theology." *Dialog* 38 (Fall): 289-293.
- . 2000. *The Problem of God in Modern Thought*. Grand Rapids: Eerdmans.
- Clayton, Philip; and Steve Knapp. 1993. "Ethics and Rationality." *American Philosophical Quarterly* (April): 97-107.

- Comte, Auguste. 1988. *Cours de philosophie positive*. Trans. as *Introduction to Positive Philosophy*, ed. Frederick Ferré. Indianapolis: Hackett Publishing Company.
- Davidson, Donald. 1993. "Mental Events." In *Essays on Actions and Events*. Oxford: Clarendon.
- . 1995. "Thinking Causes." In *Mental Causation*, ed. John Heil and Alfred Mele. Oxford: Clarendon Press.
- Dilthey, Wilhelm. 1989. *Introduction to the Human Sciences*. Ed. Rudolf Makkreel and Frithjof Rodi. Princeton: Princeton University Press
- . 1996. *Hermeneutics and the Study of History*. Ed. Rudolf Makkreel and Frithjof Rodi. Princeton: Princeton University Press.
- Earley, Sr., Joseph E. 1998. "How Constrained is the Origin of Coherence in Far-from-equilibrium Chemical Systems?," paper delivered at the Second Conference on the Philosophy of Chemistry, Cambridge University, August 3-7.
- Eccles, Sir John. 1989. *Evolution of the Brain: Creation of the Self*. New York: Routledge.
- . 1994. *How the Self Controls its Brain*. New York: Springer-Verlag.
- Edelman, Gerald M. 1992. *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York: Basic Books.
- Fodor, Jerry. 1981. "Special Sciences." In *Readings in the Philosophy of Psychology*, ed. Ned Block. Cambridge: MIT Press.
- . 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press
- . 1994. *The Elm and the Expert: Mentalsese and its Semantics*. Cambridge: MIT Press.
- Gadamer, Hans-Georg. 1975. *Truth and Method*. Ed. Garrett Barden and John Cumming. New York: Seabury Press.
- Giddens, Anthony. 1976. *New Rules of Sociological Method: A Positive Critique of Interpretive Sociologies*. London: Hutchinson.

- Grossman, M.; F. Payer; K. Onishi; M. D'Esposito; D. Morrison; A. Sadek; and A. Alavi. 1998. "Language Comprehension and Regional Cerebral Defects in Frontotemporal Degeneration and Alzheimer's Disease." *Neurology* 50 (January): 157-163.
- Hare, R. M. 1984. "Supervenience," *Aristotelian Society Supplementary Volume* 58: 1-16.
- Hempel, Carl. 1965. "Typological Methods in the Natural and the Social Sciences." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, ed. Carl Hempel. New York: Free Press.
- Henrik von Wright, Georg. 1971. *Explanation and Understanding*. Ithaca: Cornell Univ. Press.
- Hodos, William and Ann B. Butler. 1997. "Evolution of Sensory Pathways in Vertebrates." *Brain, Behavior, and Evolution* 50: 189-197.
- Kim, Jaegwon. 1993. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.
- . 1994. "The Myth of Nonreductive Materialism." In *The Mind-Body Problem*, ed. Richard Warner and Tadeusz Szubka. Oxford: Blackwell. 242-260.
- . 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge: MIT Press.
- Lenhoff, Howard M.; Paul Wang; Frank Greenberg; and Ursula Bellugi. 1997. "Williams Syndrome and the Brain." *Scientific American* (December): 68-73.
- Lycan, William. 1996. *Consciousness and Experience*. Cambridge: MIT Press.
- Maturana, Humberto; and Francisco Varela. 1992. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Trans. by Robert Paolucci. Rev. ed. New York: Random House.
- Murphy, Nancey. 1990. *Theology in the Age of Scientific Reasoning*. New York: Cornell University.
- Nagel, Ernest. 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation*. London: Routledge and Kegan Paul.

- O'Connor, Timothy. 1994. "Emergent Properties." *American Philosophical Quarterly*.
- Peacocke, Arthur. 1993. *Theology for a Scientific Age: Being and Becoming—Natural, Divine, and Human*. Enlarged ed. Minneapolis: Fortress.
- . 1999. "The Sound of Sheer Silence: How Does God Communicate with Humanity." In *Neuroscience and the Person: Scientific Perspectives on Divine Action*, ed. Robert Russell, Nancey Murphy, Theo Meyering and Michael Arbib. Vatican City State: Vatican Observatory Publications.
- Plantinga, Alvin; and N. Wolterstorff, eds. 1983. *Faith and Rationality*. Notre Dame: Notre Dame University Press.
- Penrose, Roger. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Penguin Books.
- Putnam, Hilary. 1975. "Philosophy and our Mental Life." In Putnam, *Mind, Language and Reality, Philosophical Papers*, vol. 2. Cambridge: Cambridge Univ. Press. 291-303.
- . 1981. *Reason, Truth, and History*. New York: Cambridge University Press.
- Richardson, Mark; and Wesley Wildman, eds. 1996. *Religion and Science: History, Method, Dialogue*. London: Routledge. Part II.
- Russell, Robert John; Nancey Murphy; Theo C. Meyering; and Michael Arbib, eds. 1999. *Neuroscience and the Person: Scientific Perspectives on Divine Action*. Vatican City State: Vatican Observatory Publications.
- Smith, Brian Cantwell. 1996. *On the Origin of Objects*. Cambridge, Ma.: MIT Press.
- . Unpublished paper. "God, approximately."
- Sperry, Roger W. 1983. *Science and Moral Priority: Merging Mind, Brain, and Human Values*. New York: Columbia University Press.
- Tormos, J.M.; C. Canete; F. Tarazona; M.D. Catala; A.P. Pascual; and A. Pascual-Leone. 1997. "Lateralized Effects of Self-Induced Sadness and Happiness on Corticospinal Excitability." *Neurology* 49 (August): 487-491.

Tracy, Thomas. 1999. "Evil, Human Freedom, and Divine Grace." In *Divine Action*, ed. F. Michael McLain and Mark Richardson. Lanham, Md.: University Press of America.

Weiskrantz, L. 1987. "Neuropsychology and the nature of consciousness." In *Mindwaves*, ed. C. Blakemore and S. Greenfield. Oxford: Blackwell.

Windelband, Wilhelm. 1912. *Encyclopädie der philosophischen Wissenschaften, in Verbindung mit Wilhelm Windelband*. Ed. Arnold Ruge. Tübingen: J. C. B. Mohr.