Philosophy of Mind

ABSTRACT: Contemporary philosophy of mind is largely an attempt to reconcile our scientific views of the world with our strongest held beliefs about ourselves. This project is of course relative to scientific images; in the previous century, those that achieved the most sustained philosophical reflection arose from logical positivism, neuroscience, and computability theory. After recapitulating this dialectic, I conclude by presenting forty-six open problems for the philosophy of mind.

1. Introduction- The Cartesian Tradition

Contemporary philosophy of mind is a series of footnotes to René Descartes, who was the first great thinker to contemplate reconciling modern scientific views of the world with received wisdom about ourselves. Famously, Descartes ended up holding this task to be impossible. As a result, he tried to save the new science by arguing that the lack of such reconciliation showed science proper to be essentially incomplete. Paradoxically, for Descartes, the success of modern science would thus reintroduce a mystical, dualistic, broadly Platonistic metaphysics. For the in-principle indescribability of the mind by science forces us to admit that reality bifurcates into two realms: matter, which can be accurately modeled by science, and mind, which will forever remain mysterious.

As we shall see, parts of this dialectical movement were reiterated over and over again in twentieth century philosophy of mind. And this is as it should be, minimally because the twentieth century was a period of such immense scientific and technological innovation. Every time our (in the terminology of Wilfrid Sellars) "scientific image" changes substantially, philosophical reflection on the possibility of reconciling the scientific and "manifest" images must also change. If the advance of science and technology was perhaps the overarching meta-narrative of the twentieth century, then coping with such advance is the overarching meta-narrative of twentieth century philosophy - and, following Descartes' musings - analytical philosophy of mind has been at the forefront of this process.

As Noam Chomsky (1966) persuasively shows, Descartes' strongest demonstrations for his dualism involve people's employment of language, which Descartes argued to be essentially non-mechanical. The *Discourse on Method*, contains the following interesting passages.

> understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for instance, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on. But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. (Descartes, (1985, p. 140))

But then, if the human mind can do something no machine can, and if the universe as treated by science is a machine, it follows that the human mind exists outside of the realm treated by science.

As the history of philosophy in the intervening three centuries would show, this argument has wide application. Consider the mind's three paradigmatic tasks. The mind represents the world through perception, beliefs, desires, and feelings. The mind reasons about which beliefs and desires it is appropriate to have as well as about how best these beliefs and desires should issue in action. The mind initiates, and is responsible for action. For all of these activities, we can ask, "how do we do that?" Part of Descartes' genius is that we can never again ask this question without prefixing, "given what the natural sciences tell us about the universe. . ."

In what follows, I will sketch three developments in, respectively, the philosophy of science, biology, and computability theory, showing how each fundamentally altered our understanding of ourselves. The first is logical positivism, not a scientific theory *per se*, but rather a scientifically minded philosophical movement that had tremendous influence for the methodology and subject matter of the social sciences. At its most extreme, this tendency yields an identification of the mind with dispositions to behavior. The second development concerns advances in the brain sciences, in particular success in localizing mental phenomena in the brain. At its most extreme, this tendency yields an

identification of the mind with the brain. The third concerns advances in the theory of computation and computer science. Here the mind can be identified with computational machinery. Of course, the lion's share of each story is careful philosophical analysis of what it means to make such identifications.

After showing the main different ways these identifications have been worked out and criticized, I will conclude with a discussion of important open questions in the philosophy of mind, ones to which we can hope that this century's philosophers will provide new insight. In presenting these, I will of necessity discuss other parts of the scientific image, as well as other philosophical areas and traditions. This will involve touching upon what might be called "non-standard" approaches to Descartes' problem. Philosophies of mind coming out of the three scientific developments all, for the most part, involve: (a) largely accepting the manifest image as is, (b) largely accepting the scientific image as is, and (c) being confident of the possibility of philosophical resolution of the tensions between the two. Following Descartes himself, many contemporary philosophers of mind think that the problem should be dissolved rather than solved, by giving up some combination of (a), (b), and (c).

For now, suppose "standard" solutions are in the offering. That is, suppose that after centuries of sustained labors of some of the best minds of each generation, we could resolve the tensions between the scientific and manifest images. Even so, we shall never be able to recover our pre-Cartesian naivete. This is all for the good, for philosophy commands

> We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive at where we started
> And know the place for the first time. (Elliot, (1950, p. 145))

There is no going back to before Descartes.

2. Logical Positivism

Following Immanuel Kant, early logical positivists such as Rudolf Carnap, Hans Schlick, and Otto Neurath divided propositions into either analytic or synthetic, and either *a priori* or *a posteriori*. Through a subtle recharacterization of these concepts, they hoped to dissolve the philosophical impasses created by neo-Kantian and neo-Hegelian philosophical schools of the early 20th century.

The positivists defined analyticity in the following manner.

> A proposition *P* is *analytic* if, and only if, *P* is true (or false) solely in virtue of the meanings of the words in *P* (and the principles governing the way those words are put together to make *P*).

One way to test for a sentence's being analytically true is to consider if it is possible for the sentence to be false without changing the meaning of any of the words in the sentence. If it is not possible to do this, then we have good evidence that the sentence is analytically true. Prima facie examples include conceptual necessities such as, "All bachelors are unmarried," and logical truths such as, "Either Jon smokes, or Jon doesn't smoke."

Conversely, syntheticity can be defined in the following manner.

> A proposition *P* is *synthetic* if, and only if, *P* is true (or false) in virtue of the meanings of the words in *P* (and the principles governing the way those words are put together to make *P*) and the way the world is.

A good test to see whether a sentence is synthetic is to see if it can be made false by the world being different than the way it is when the meaning of the words is held fixed. Consider the sentence, "Bill Clinton was President." This could be false even when the words meant the same as they do now, but the world was such that the elder Bush had gotten re-elected.

Importantly, the analytic/synthetic distinction is a semantic distinction; it only concerns what makes sentences true or false. The *a posteriori*/*a priori* distinction is an epistemological distinction; it concerns what justifies our knowledge of the truth of certain sentences.

> A proposition *P* is *a priori* if, and only if, *P* can't be proven (or disproven) from experience.

> A proposition *P* is *a posteriori* if, and only if, *P* can be proven (or disproven) from experience.

As Kant realized, a major source of philosophical perplexity at least since Plato has involved the difficulty in seeing how sensory experience could justify mathematical claims (for example, those involving infinite totalities or geometrically perfect shapes) or ethical claims (about what we ought or ought not do). The fact that such propositions also seem to make substantive claims about the world led Kant to characterize mathematic and ethical propositions as synthetic *a priori*. But then there are claims that are true of the world yet such that our sensory experience of the world does not justify belief in them.

The Platonist response to this problem is similar to (and indeed the model for) Descartes' response to his problem. If we have knowledge of the world that is not justified by sensory experience, then there must be an part of the world that we commune with that cannot be experienced by our sensory facilities, the realm of Plato's forms. The Kantian response is to locate justification for synthetic *a priori* judgments not in knowledge of the world of experience, but rather in how our minds process experience. The Hegelian

response generalizes the Kantian by postulating that the world itself can be thought of in terms of the properties we attribute to minds.

Logical positivism was a fundamental rethinking of this Platonistic/Kantian/Hegelian dialectic. Instead of trying to solve the problem of synthetic *a priori* truths, the logical positivists attempted to dissolve it, mainly by using modern logic to characterize analyticity and syntheticity such that the class of synthetic *a priori* propositions would be empty. For logical positivists, analytic *a priori* claims are those that follow by logic from meaning characterizing definitions. Synthetic *a posteriori* claims are those that play a role in helping us predict sensory experience. In this manner, *pace* Kant, a philosophy of mathematics was developed that would render math analytic *a priori*, and a philosophy of science was developed that would render claims about space, time, and causality synthetic *a posteriori*.

Formal logic played a two-fold role. First, it was the manner in which one derived mathematical theorems from the meaning characterizing axioms, and second, it was the manner in which predictions were to follow from scientific theories. Though the use of logic thus went in the same direction, there is an important asymmetry that must be noted. In the case of mathematics a logical proof is part of the justification for the theorems. In the case of synthetic *a posteriori* claims, it is just the opposite. Being able to derive correct predictions from a theory end up justifying that theory. In math the order of justification goes from axioms to theorems via logic. In science the order of justification goes from predicted observation statements to the theory, in the reverse order of the logical deductions.

Recent scholarship on logical positivism (e.g. (Coffa, 1991) and (Friedman, 1999)) has stressed how much of Kant's philosophy was retained by the positivists. Although they attempted to overthrow the hegemony of the synthetic *a priori* in the manner here sketched, they reaffirmed Kant's verificationism and inferentialist (see (Brandom, 2001)) model of word meaning and concepts, and it is these strains that give rise to behaviorism.

P.F. Strawson describes Kant's verificationism in this manner

> [Kant's principle of significance] is the principle that there can be no legitimate, or even meaningful, employment of ideas or concepts which does not relate them to empirical or experiential conditions of their application. If we wish to use a concept in a certain way, but are unable to specify the kind of experience-situation to which the concept, used in that way, would apply, then we are not really envisaging any legitimate use of that concept at all. In so using it, we shall not merely be saying what we do not know; we shall not really know what we are saying. (Strawson, (1966, p. 16))

From the discussion above of the logical positivist's characterization of synthetic *a posteriori* and analytic *a priori* beliefs, it should be clear how the positivists embraced this principle. Analytic *a priori* claims are verified by logical proof from meaning characterizing axioms and synthetic *a posteriori* claims are verified by being part of a

theory from which one can prove true observation statements. In this manner, evidential and inferential roles were what constituted the meanings of parts of language. Thus were they both verificationists and inferentialists.

Positivists used these Kantian positions in both constructive and critical ways. Constructively, they laid the foundations for contemporary philosophy of mathematics, logic, the philosophy of science, and the statistical methodologies of the social sciences. Critically, they used this verificationism to criticize pseudo-science and (again, like Kant) metaphysics, which they saw as contradicting verificationism. The three varieties of behaviorism dominant in the twentieth century (analytic behaviorism, methodological behaviorism, and interpretavism) all combine these constructive and critical aspects.

a. Analytical Behaviorism

In his 1935 presentation, "The Logical Analysis of Psychology," Carl Hempel considers the philosophy of mind appropriate to logical positivism. Hempel first notes that in science there are often equivalent ways to state the same information, for example one can give temperature using a centigrade or Fahrenheit scale. For that matter, one can give temperature in units of measurement from alcohol thermometers or mercury thermometers, among many other such kinds of apparatus. One can translate a sentence about temperature from one idiom to the other, and each sentence will express the same information about the temperature in question, which he holds to be "an abbreviated expression of the fact that all of its test sentences are verified" (Hempel, (1980, 18)).

Hempel then argues that the example of temperature is completely analogous to the example of mental properties such as being in pain. In the case of having a toothache, he gives the following,

> (a) Paul weeps and makes gestures of such and such kinds.
>
> (b) At the question "What is the matter?," Paul utters the words "I have a toothache."
>
> (c) Closer examination reveals a decayed tooth with exposed pulp.
>
> (d) Paul's blood pressure, digestive processes, the speed of his reactions, show such and such changes.
>
> (e) Such and such process occur in Paul's central nervous system. (Hempel, (1980, p. 17))

While Hempel admits that this list would have to be greatly expanded, he takes it to show "the essential point, namely, that all the circumstances which verify this psychological statement are expressed by physical test sentences." (Hempel, (1980, p. 17)) Thus, for Hempel, to attribute a mental property to someone is just to attribute characteristic behavior and dispositions to behave to that person. To be charitable, we must note that

for analytical behaviorists such as Hempel, measurable bodily states such as blood pressure count as part of behavior.

i. Some Reasons for Analytical Behaviorism

In addition to analytical behaviorism's clear fit with the logical positivist *zeitgeist*, Hempel did present an independent argument for the view, which Jaegwon Kim reconstructs in the following manner.

> (1) The content, or meaning, of any meaningful statement is exhausted by the conditions that must be verified to obtain if we are to consider that statement true (we may call them the "verification conditions" of the statement).
>
> (2) If a statement is to have an intersubjective content, that is, meaning that can be shared by different persons, its verification conditions must be publicly observable.
>
> (3) Only behavioral and physical phenomena are publicly observable.
>
> (4) Therefore, the content of any meaningful psychological statement must be specifiable by statements of publicly observable verification conditions, that is, statements describing appropriate behavioral and physical conditions that must hold if and only if the psychological statement is to count as a true. (Kim, (1998, p. 30))

Here we see how the positivist's Kantian combination of inferentialism and verificationism can give rise to a philosophy of mind.

ii. Objections

In "Brains and Behavior" Hilary Putnam delivered this philosophy's deathblow, arguing that any attempted behaviorist definition of mental states will be viciously circular and that the philosophy of science motivating it is mistaken.

Putnam first notes that his attack will work against a position logically weaker than analytical behaviorism, one comprised of the two theses:

> (a) That there exist entailments between mind-statements and behavior-statements; entailments that are not, perhaps, analytic in the way in which 'All bachelors are unmarried' is analytic, but that nevertheless follow (in some sense) from the meaning of mind words. I shall call these analytic entailments
>
> (b) That these entailments may not provide an actual translation of 'mind talk' into 'behavior talk'. . . but that this is true for such superficial reasons as the greater ambiguity of mind talk, as compared with the relatively greater specificity of overt behavior talk. (Putnam (1980a, p. 25))

Since analytical behaviorism entails (a) and (b) Putnam's critique of (a) and (b) will undermine analytical behaviorism.

1) The Circularity Objection

Putnam's argument is by analogy. He first draws our attention to the fact that when a virus origin was discovered for polio, doctors said that many cases where all of the symptoms were present, but the virus was not, were not actually cases of polio. Thus we should reject the view that "polio" means "the simultaneous presence of such and such symptoms." Rather, we should take it to mean "that disease which is normally responsible for some or all of the following symptoms. . ." Analogously, as Putnam will argue, mental states are responsible for behavior, not mere dispositions to behave.

As Putnam notes, an analytical behaviorist would respond by claiming that the meaning of the word "polio" changed when scientists discovered the virus and began to allow the presence or absence of the virus to be the deciding factor in whether or not someone has polio. But this is odd, because then if a doctor were talking prior to the discovery and said, "I believe polio is a virus" they would then be saying something false.

> ". . .and if a doctor even said (and many did) 'I believe this may not be a case of polio', knowing that all of the textbook symptoms were present, that doctor must have been contradicting himself (even if we, today, would say that he is right)." (ibid., 27)

We can state this a little more formally as:

> (1) If "polio" means "the simultaneous presence of such and such symptoms," then a doctor who says "Sam has all of the symptoms associated with polio but (since he doesn't have the virus) does not have polio" is either (i) saying something false, or (ii) changing the meaning of the word "polio."
>
> (2) (i) is clearly wrong as Doctors do say this and say it truly.
>
> (3) On reflection we can see that (ii) is wrong. Suppose that two doctors disagreed about whether or not someone could have all of the polio symptoms and have polio. Their dispute could not be solved by consulting a dictionary. That is, it is not a dispute about the proper meaning of the word "polio."
>
> (4) Rather, doctors take themselves to be talking about the same thing, but disagreeing about the properties of that thing. If "polio" is understood to mean "whatever is responsible for polio symptoms" then such disagreement makes sense.

Thus, if you think of polio as the disease normally responsible for the symptoms, you can consistently: (a) speak of discovering the virus that causes the polio symptoms, (b) discover that someone lacks polio even though they have all of the symptoms, and (c) discover that someone has polio even though they have none of the symptoms.

Therefore, statements about diseases such as polio are not translatable into talk about symptoms, simply because causes are not logical constructions out of their effects.

Then Putnam goes on to argue that "pain" behaves the way he takes "polio" to, ("just as before causes (pains) are not logical constructions out of their effects (behavior)"). He first states that:

> It is possible for someone to be in pain and not manifest any of the normal pain behavior.

He argues for this by positing the possible world existing of superspartans, who lack all possible "behavior" we associate with pain. Again, the cause/effect hypothesis explains this well. With any causal process of sufficient complexity you can imagine possible worlds where things interfere such that the causes can take place without the effects.

He also considers beings who experience pain because of very different causes than what cause us to experience pain (X rays cause them pain so they can be called "X worlders"). Then, if these creatures were also super-spartans, they would be such that they were in pain while the stimulus-response pairs that characterize their pain are completely different from ours.

Now he can argue that both strands of logical behaviorism are false, (a) There are no entailments between pain statements and behavior statements, because pain only causes certain kinds of behavior in combination with one's "beliefs, desires, ideological attitudes, and so forth." (30) So the statement "x is in pain" itself entails nothing about behavior, but then (b) there clearly won't be a translation from pain talk into behavior talk which preserves meaning.

Another way to view Putnam's criticism is to note that it entails that mental talk cannot be defined purely in terms of behavioral talk, because possession of a mental state depends upon possession of other mental states too. But then any attempted behaviorist definition will be viciously circular. This kind of criticism was originally made, in a different context, in Roderick Chisholm's *Perceiving*. Chisholm argued that the success of the attempt to define everyday objects in terms of sense data requires specifying the relevant sets of sense data without referencing any other everyday objects. In the context of analytical behaviorism, Putnam turns this criticism on its head. Again, if one had to mention, for example, the belief that suffering is noble in one's set of dispositions corresponding to being in pain, then one would not have defined mental talk in terms of behavioral conditions, but rather in terms of behavioral conditions and mental states. Though this critique does undermine analytical behaviorism, when we discuss Ramsey Sentence Functionalism (section 4b.), we shall see that there are senses in which it might be answered from a broadly behaviorist standpoint.

<u>2) The Explanation Objection</u>

The analytical behaviorist might appeal to the logical positivist's verificationist mantra (that if something is untestable then it is meaningless) to argue that Putnam's possible worlds are irrelevant. The logical positivist might say we would have no testable means to distinguish X worlders from non-X worlders who really don't know what pain is. The thought that we could look at their brains would be equally problematic for the positivist; how do we know that superspartans' brain states correspond to their phenomenal states in the same way that ours do?

Putnam argues that the notion of "testable" a sound philosophy of science employs would be such that the differences between a superspartan and people are testable, since considerations of simplicity, elegance, fecundity, fit with other theories, etc. are part of how we test scientific theories. We can think of other scenarios where we might get evidence for the claim, for example say a language of thought exists in common to us and the superspartans, and that this language can be read with a V decoder. If we got "I'm in pain" readings from the V decoder when pointed at the superspartans we might have good reasons to think they were in pain because "no other likely explanation readily suggests itself." (ibid., 33) Putnam argues that this is a perfectly fine sense of verifiability that we must use in the sciences constantly to determine, for example, what is going on in the sun.

By looking at why Hempel's argument originally failed we can understand Putnam's critique better. Again, consider Kim's presentation of Hempel's argument.

> (1) The content, or meaning, of any meaningful statement is exhausted by the conditions that must be verified to obtain if we are to consider that statement true (we may call them the "verification conditions" of the statement).
>
> (2) If a statement is to have an intersubjective content, that is, meaning that can be shared by different persons, its verification conditions must be publicly observable.
>
> (3) Only behavioral and physical phenomena are publicly observable.
>
> (4) Therefore, the content of any meaningful psychological statement must be specifiable by statements of publicly observable verification conditions, that is, statements describing appropriate behavioral and physical conditions that must hold if and only if the psychological statement is to count as a true. (Kim, (1998, p. 30))

Putnam can be understood as arguing that any notion of verification conditions plausible enough to render natural science meaningful will be such that premise (3) is false, and mental phenomena will be seen to be in fact publicly observable. At least they are no less observable than electrons or other entities in which the sciences must traffic. Consider an analogous argument.

> (1') The content, or meaning, of any meaningful statement is exhausted by the conditions that must be verified to obtain if we are to consider that statement true (we may call them the "verification conditions" of the statement).

(2') If a statement is to have an intersubjective content, that is, meaning that can be shared by different persons, its verification conditions must be publicly observable.

(3') Only macro-world phenomena such as readouts on instruments are publicly observable.

(4') Therefore, the content of any meaningful statement of physics that refers to theoretical entities that are not directly observable must be specifiable by statements of publicly observable verification conditions, that is, statements describing appropriate macro world physical phenomena such as instrument readouts that must hold if and only if the physics statement is to count as a true.

Putnam's point is that, to the extent that premises (1') and (2') are plausible, then the "publicly observable" verification conditions of sentences about electrons and the such include broader notions of theoretical economy. For example, that a physical theory that is easier for us to use involves reference to electrons is some verification of the claim that electrons exist. But then statements about mental states are verifiable, and thus there is no need to think that the translations in (4') are needed.

This undermining of the bad philosophy of science underlying analytical behaviorism proves to be decisive. As we shall see (section 2c.), a key impetus for the development of Daniel Dennett's interpretivism, which should be thought of as the natural logical positivist philosophy of mind that results when good philosophy of science is used.

b. Methodological Behaviorism

Logical positivism was one of the most influential philosophies of the previous century not just because of how it fundamentally changed the way philosophers see the world, but because it radically changed how non-philosophers live their lives. In particular, the social sciences have never been the same. Again, following Kant, the critical aspect of Positivism entailed that if sentences did not have verification conditions then they were nonsense. While this was used by philosophers to undermine the neo-Hegelian and neo-Kantian schools then controlling academic philosophy departments in the English speaking world, it was also used in the social sciences to encourage the development of experimental methods that lent themselves to quantitative modeling and double-blind verification of results.

This precisification of the discipline of psychology gave rise to a revolutionary movement which can be called "Methodological Behaviorism," the position that psychology should only talk about behavior, never mental states. It is very important to realize that behaviorist psychology differed from philosophical analytical behaviorism in two respects, (1) the conception of "behavior" was much narrower, and did not include observable facts about an agent's physiology, and (2) no claim whatsoever about the translation of "mind talk" into "behavior talk" was ever made. This form of behaviorism is completely unscathed by Chisholm and Putnam's critiques.

i. Reasons for Analytical Behaviorism

B.F. Skinner, the prophet of methodological behaviorism, gave the canonical arguments in the influential *Science and Human Behavior*. Perhaps the most influential is the following:

(1) The point of psychology is to control and predict behavior.

(2) Mention of inner causes such as human physiology and states of mind is completely unnecessary in devising theories that allow us to control and predict behavior.

(3) Moreover, research involving operant conditioning [defined below] will allow us to devise theories where human behavior is solely a function of stimulus. Such theories will allow us to control and predict human behavior very well.

Thus, methodological behaviorists: (a) refuse to talk about inner mechanisms, (b) do experiments involving operant conditioning, and (c) attempt to characterize behavior as a function of stimulatory inputs.

The first and third part of Skinner's argument are hard to evaluate philosophically. In particular, the third claim is a promissory note, so we can merely ask whether or not such theories have been developed, and, if not, try to determine why not. Luckily, the young Noam Chomsky turned his formidable intellect to this task in his "Review of *Verbal Behavior*," which we will discuss below.

Skinner himself gave a few arguments in *Science and Human Behavior* for the second claim, which we can evaluate easily with the benefit of hindsight. His first argument depended upon the state of psychology at the time. It can be represented as:

(1) Direct information about inner mechanisms is seldom, if ever available.

(2) So such information is not helpful in predicting behavior.

(3) We have no way of currently directly altering inner mechanisms.

(4) So information about those mechanisms is not helpful in controlling behavior.

(5) Thus, since the point of scientific psychology is to predict and control behavior, information about inner mechanisms is irrelevant.

If by "inner mechanisms" one means to be talking about the human brain, then as we will show in our discussion of the identity theory (section 3 of this article), the psycho-pharmacological revolution has proven Skinner's pessimism here completely unfounded.

Sometimes Skinner makes a much stronger argument though:

> (1) Human action is the result of: (a) an operation performed upon the organism from without, (b) an inner condition.
>
> (2) Thus, the external stimulus brings about some change in inner condition which brings about the human action.
>
> (3) But then, "unless there is a weak spot in our causal chain so that the second link is not lawfully determined by the first, or the third by the second, then the first and third links must be lawfully related." (Skinner, (1980, p. 42))
>
> (4) But then, we may simply "avoid many tiresome and exhausting digressions by examining the third link as a function of the first." (ibid.)
>
> (5) Thus, inner states are "not relevant in functional analysis." (ibid.)

In our discussion of Ramsey sentence functionalism (section 4b. of this article), we will show that there is a logical sense in which this argument can be shown to be valid. However, as we shall show, problems with Ramsey sentence functionalism ultimately affect the soundness of this argument as well.

ii. Objections

Both of the main objections to analytical behaviorism stem from the early work of Noam Chomsky. Prior to making them, we need to be clear about a few behaviorist terms of art.

> *operant*- a class of responses determined by the similarity of the consequences of those responses. In operant conditioning the operant is defined by the property upon which reinforcement depends (i.e. the activity of pressing a bar).
>
> *operant conditioning*- the process of changing frequency with which an operant occurs as a result of reinforcement (i.e. the rat is taught to press a bar)
>
> *probability of response*- the frequency of response to a stimulus in standard conditions, sometimes this is called the strength of the operant (i.e. how often and how many times the rat will continue to press the bar after food is no longer coming out as a result).
>
> *response differentiation*- when a subject's response is changed as a result of the schedule of the reinforcement changing (i.e. the rat is only rewarded when he pushes the bar for a certain duration, and then starts to press the bar just for that duration).
>
> *stimulus discrimination*- when a subject responds not merely to reinforcement, but stimulus coupled with the reinforcement (i.e. a light must be flashing for pressing the bar to cause food to be released, and the rat starts to only press the bar when the light

flashes). This new stimulus is called the *discriminated operant*. If it itself becomes reinforcing because it is repeatedly associated with reinforcers it becomes a *secondary reinforcer*.

As an example of secondary reinforcement, one can tell a story about why people want social status in terms of social status repeatedly being associated with things that are innately reinforcing. This much is common sense. Aristotle noted essentially the broader point in his discussion of child rearing in the *Nichomachean Ethics*. However, Chomsky argued that such common sense is not a sufficient foundation for psychology.

1) The Equivocation Objection

In the majority of "A Review of B.F. Skinner's Verbal Behavior," Chomsky attacks the pretense that Skinner's book *Verbal Behavior* provides any evidence at all for the third line of Skinner's Überargument: "Moreover, research involving operant conditioning will allow us to devise theories where human behavior is solely a function of stimulus. Such theories will allow us to control and predict human behavior very well." In the critique, Chomsky sets forth the following sort of dilemma.

(1) Either technical words such as *stimulus*, *response*, *reinforcement* have precise technical meanings or not.

(2) In the rat labs *stimulus* refers not merely to all physical events to which the organism is capable of reacting, but rather to events to which the organism really does react, and *response* (or *operant*) is used to refer not to any behavior whatsoever, but only ones that can be connected to stimuli in lawful ways.

(3) If we use these terms with the same meaning they have in the rat labs, then there has been no demonstration whatsoever by Skinner or anyone that linguistic learning (or perhaps any mental state for that matter) can be explained in terms of operant conditioning (no functional analysis, nor any data suggesting one, has ever been given).

(4) If we use the terms in the more broad senses, then the scientist must conclude that the relevant behavior is not lawful, as no functional analysis is possible unless the stimuli and responses are specified.

(5) Hence, "the psychologist either must admit that behavior is not lawful, or must restrict his attention to those highly limited areas in which it is lawful. (e.g. with adequate controls, bar-pressing in rats; lawfulness of the observed behavior provides, for Skinner, an implicit definition of a good experiment)" (Chomsky, (1980, p. 51)) Thus, "if we take [Skinner's] terms in their literal meaning, the description covers almost no aspect of verbal behavior, and if we take them metaphorically, the description offers no improvement over various traditional formulations." (ibid., p. 57)

Of Skinner's book, Chomsky writes:

> Skinner does not consistently adopt either course. He utilizes the experimental results as evidence for the scientific character of his system of behavior, and anologic guesses (formulated in terms of a metaphoric extension of the technical vocabulary of the laboratory) as evidence for its scope. This creates the illusion of a rigorous scientific theory with a very broad scope, although in fact the terms used in the description of real-life and of laboratory behavior may be mere homonyms, with at most a vague similarity of meaning. (ibid., p. 51)

Then in most of the article Chomsky does establish that the "technical" terms in Skinner's book are used so broadly that Skinner's behaviorist explanation is not really an explanation at all, but rather a statement of platitudes that merely obfuscate real issues.

> A typical example of *stimulus control* for Skinner would be the response to a piece of music with the utterance *Mozart* or to a painting with the response *Dutch*. These responses are asserted to be "under the control of extremely subtle properties" of the physical object or event (108). Suppose instead of saying *Dutch* we had said *Clashes with the wallpaper, I thought you liked abstract work, Never saw it before, Tilted, Hanging too low, Beautiful, Hideous, Remember our camping trip last summer?*, or whatever else might come into our minds when looking at a picture (in Skinnerian translation, whatever other responses exist in sufficient strength). Skinner could only say that each of these responses is under the control of some other stinmulus property of the physical object. If we look at a red chair and say *red*, the response is under the control of the stimulus *redness*, if we say *chair*, it is under the control of the collection of properties (for Skinner, the object) *chairness* (110) and similarly for any other response. This device is as simple as it is empty. Since properties are free for the asking (we have as many of them as we have nonsynonymous descriptive expressions in our language, whatever this means exactly), we can account for a wide class of responses in terms of Skinnerian functional analysis by identifying the *controlling stimuli*. But the word *stimulus* has lost all objectivity in this usage. Stimuli are no longer part of the outside physical world; they are driven back into the organism. We identify the stimulus when we hear the response. It is clear from such examples, which abound, that the talk of *stimulus control* simply disguises a complete retreat to mentalistic psychology. We cannot predict verbal behavior in terms of the stimuli in the speaker's environment, since we do not know what the current stimuli are until he responds. (ibid., p. 52)

In the vast majority of the article Chomsky devastatingly shows how *all* of the technical behaviorist vocabulary defined above is used in this manner in Skinner's *Verbal Behavior*.

<u>2) The Poverty of Stimulus</u>

In the last section of his review, Chomsky suggests what is usually referred to as "Chomsky's poverty of stimulus argument." This argument can be presented in the following manner.

> (1) Assume that Skinner is right, and that verbal behavior can be explained by a theory that restricts itself to presenting such behavior as the output of a function whose input is sensory stimulus, a function whose "computation" only involves attributing to the speaker general learning strategies of the type amenable to operant conditioning.
>
> (2) Part of our verbal behavior is the ability to tell of arbitrary sentences whether or not they are grammatical in various respects.
>
> (3) Thus, if Skinner is right, our ability to tell of arbitrary sentences whether or not they are grammatical in various respects should be able to be presented as the output of a function with sensory stimulus as input and which involves operant conditioning in getting from input to output.
>
> (4) The syntactic part of a grammar is like an extremely complex system of logic where grammatical sentences are derived by proofs in the system, and non-grammatical strings of words are not derived.
>
> (5) But when we examine the stimulus that a child is exposed to and compare how impoverished and paltry this is to how complicated even the syntactic part of a grammar is, we see that the ability to tell whether or not an arbitrary sentence is grammatical in various respects simply cannot be predicted in a theory which only refers to stimulatory input and generalized learning strategies of the type amenable to operant conditioning (or any other generalized learning mechanism).

Chomsky also suggests that behaviorist linguistics can be replaced with "Cartesian Linguistics" where the proper subject of scientific study is taken to be "what is in a speaker's head." One of the benefits of this is that it sets in bold relief the task of the psycholinguist who wants to explain language acquisition.

> (1) On the other hand, if we understand linguistic abilities to involve tacit knowledge of a grammar, we can (in a manner unacceptable to the behaviorist) develop sound scientific theories of language acquisition.
>
> (2) If we understanding linguistic abilities as involving tacit knowledge of a grammar, then the child's learning task is to determine what the grammar of her language is from the stimulatory input she has.
>
> (3) The seeming inexplicability of anyone's ability to do this, given the kind of stimulatory input available to the child, is not so inexplicable if the child is not following generalized learning strategies, but rather is innately highly constrained in various ways to hypothesize and test the grammar of her language.

(4) Thus, our task is to both develop the grammars that competent speakers know, and to investigate the ways in which a child's hypotheses are innately constrained. In this way we shall be able to isolate what the child brings to the learning process (both in terms of what can count as a possible grammar, as well as how stimulatory inputs allow one to pick one of the possible grammars).

Chomsky himself often tries to use these concerns to motivate specific approaches to syntax, and to criticize approaches which he thinks are not sufficiently "psychologically real." We shall not be able to consider such arguments, because fully understanding and adjudicating them requires understanding different syntactic frameworks. The important point is that the vast majority of linguists agree that the early behaviorist strictures on scientific explanations would render scientific explanation of language impossible, for the very reasons that Chomsky pointed out. Moreover, the vast majority of contemporary syntax is "Chomskyan" in the sense that all dominant theories are descendents of Chomsky's early *Syntactic Structures* era theory.

iii. A New Behaviorism?

While Chomsky's early criticisms of behaviorist strictures were an important correction, one should not conclude too much from them. It is now respectable to theorize about innate capacities and inner mechanisms. But it is quite a stretch from this to Chomsky's current theory of language acquisition, and the way this theory is supposed to constrain syntax and semantics. For Chomsky (1995), there is a "universal grammar" describable by the linguist, invariant over all languages, and known by the infant prior to language learning. A grammar for a natural language is, in some manner, generated by setting parameters on the universal grammar. Such parameter setting is supposed to correspond to stages of language acquisition. It is even more of a stretch from these *a priori* restrictions on grammar and theories of learning, to the view that Chomsky (1996) shares with Jerry Fodor concerning people's *a priori* knowledge, for example that people have innate concepts of carburetors.

These hypercognitivist views have come under scathing attack recently in linguistics, psychology, and philosophy. For example, Fiona Cowie (1999) carefully examines and undermines Chomsky and others' arguments for innate knowledge. Likewise, in the intervening years since Chomsky's review of Skinner, connectionist approaches to artificial intelligence have demonstrated to many people's satisfaction how generalized learning algorithms could be sufficient for learning grammar (Sun, 2001). Finally, people involved in computational linguistics have abandoned Chomsky's universal grammar as computationally unworkable (Johnson & Lappin, 1997 and 1998). Computationalists work in competing frameworks such as Head Driven Phrase Structure Grammar (Pollard & Sag, 1994) or Categorial Grammar (Carpenter, 1998).

For many contemporary cognitive scientists (e.g. Clark, 1998), the challenge now is to show how a massively distributed, analog parallel processor like the brain (or like a connectionist system) could develop digital, rule-based capacities (like Categorial Grammar). While this is non-behaviorist in the sense that it is now fair to talk about

internal mechanisms, it represents an extreme move away from what many view as Chomskyan cognitivist excesses towards a view much more Skinnerian, in so far as *a priori* knowledge is minimized and generalized learning procedures are embraced. An influential, accessible, and insightful discussion of the problems and prospects for a new behaviorism is J. Staddon's *The New Behaviorism: Mind, Mechanism, and Society*. The philosophical prophet of this new movement is Daniel Dennett, to whom we now turn.

c. Interpretivism

Of all living philosophers, Daniel Dennett has done more than anyone to rehabilitate a verificationist philosophy of mind. His views are so subtle and far-reaching that it is impossible to do justice to them in an encyclopedia article. Instead, we can paint his starting point in broad strokes, showing how a liberalized form of verificationism allows him to navigate between the Scylla of Skinnerian behaviorism and the Charibdis of Chomskyan cognitivism.

Dennett can best be presented as a neo-positivist who has wholeheartedly accepted Putnam's criticism of Hempel. Remember that Putnam argued that fundamental entities of our physical theories are not definable in terms of the observations associated with them. Rather, verifying that, say, electrons exist involves determining that a theory which says electrons exist is better than a theory which does not. Dennett ingeniously applies this kind of thinking to mental entities such as beliefs and desires.

Empirical theories are better and worse than one another in a variety of ways. One may be simpler than another, one may fit better with other theories we accept, one may be more fruitful in yielding technology than another. . . However, one thing in common to all empirical theories is that they are supposed to yield predictions that are intersubjectively testable. If one theory entails that a measuring device will read "3" and another entails that it will read "7," and the meter reliably reads "3.1," then we have good reason to prefer the former.

Dennett notes that such issues are not restricted to abstract debates in the philosophy of science. All of us are able to make extraordinarily reliable predictions about the world. If I knock a pen off the table I know it will fall. If I set my alarm clock correctly, I know it will ring the next morning. If I postpone a test, I know my students will smile. Dennett (1989) develops a typology of such predictions in terms of the most basic presuppositions we make when engaging in them. He divides these into three stances: the physical, the design, and the intentional.

We take the physical stance when the presupposition that a system evolves according to causal laws allows us to predict that system (e.g. the behavior of the falling pen). While the physical stance is ubiquitous, it is insufficient for many predictions. For example, it is not in practice possible to predict the evolution of a video game playthrough by knowing the physical state of the machine. We also need to know for what the machine was designed. When we use the assumption that a system is designed for some purpose we are engaged in the design stance.

But the design stance is again not sufficient for all of our predictive behavior. For some systems (paradigmatically people, but also machines, animals, and perhaps other systems too) correct prediction requires assuming that the system has beliefs, desires, and at least a modicum of reason. This assumption is the intentional stance.

For Dennett, just as verification of electrons proceeds by accepting an electron-positing theory as best, verification of the existence of beliefs, desires, and reason requires accepting a belief/desire/reason-positing theory as best.

At this point one should ask whether Dennett thinks that such belief-desire entities are really real, or whether they are just predictively useful fictions (in the way phlogiston and caloric were for now disconfirmed theories). Dennett (1998) argues to the extent that the intentional stance is indispensable (to the extent that we cannot make such predictions from the physical or design stance) the relevant entities (beliefs, desires, feelings, etc) should be though of as real.

While this position is clearly free from the traditional problems of behaviorism, one could fairly object that it is not a complete philosophy of mind, for all it does is tell us that belief/desire talk is indispensable. It does not tell us why such talk works. Dennett would agree with this. In fact, the main function of Dennett's theory of stances is that it slightly recasts the main tasks of the philosophy of mind. Rather than somehow grounding cherished truths about ourselves (the "manifest image") we are now only tasked with explaining why such putative truths work, why the assumption of such truths allow us to predict behavior correctly. In fact, it is consistent with this aspect of Dennett's program to hold that the manifest image is false. All one need to explain is why it works. As we shall see in our discussion of the Churchlands, who argue that it is highly probable that at least some of our manifest image is mistaken, this is an important concession.

Moreover, this is exactly where, for the Dennettian, avoiding the Charibdis of Chomskyan cognitivism comes in. It is hoped that by attending to brain science, computer science, evolutionary theory, and research into interpretation, we will be able to explain why the intentional stance works in a way that does not commit us to a Fodorian (1980) "language of thought" with implausible *a priori* concepts. For the Fodorian, beginning with the assumption that meanings are determinate and thinking at its very basis linguistic, it is (at least as Fodor and Chomsky argue) impossible to avoid such conclusions. But the Dennettian need only explain why the presupposition of determinacy and the linguistic basis of thinking allow us to predict others' behavior. This is then consistent with powerful arguments by some philosophers of language (e.g. Wilson (1982)) that meaning is far less determinate that we presuppose it to be in our normal conversational settings.

Of course Dennett's project is too big to assess in an article such as this. In addition, it is still ongoing. However, since Dennett's project requires knowledge of the relevance of

brain science and computabilty theory to the philosophy of mind, being able to assess and engage in Dennett's project presupposes knowledge of that to which we now turn.

3. Neuroscience

By the early twentieth century, areas of the brain had been mapped out that corresponded to discreet abilities involving vision, hearing, feeling, taste, smell, motor functions, sleep, emotion, memory, speech, and understanding. Much of this knowledge was arrived at by dissecting the brains of the dead. A paradigm example was Wernicke's discovery that people who habitually speak syntactically correct and semantically meaningless sentences (technical label "word salad") have usually suffered damage to the left hemisphere of their brains. (For an excellent description of this history, see (Finger, 1994).)

This kind of data leads one to begin to doubt Descartes' certainty that the mind must be separate from the brain. For example, reflecting on earlier and recent discoveries, Patricia Churchland writes:

> The degeneration of cognitive function in various dementias such as Alzheimer's disease is closely tied to the degeneration of neurons. The loss of specific functions such as capacity to feel fear or see visual motion are closely tied to defects in highly specific brain structures in both animals and humans. The shift from being awake to being asleep is characterized by highly specific changes in patterns of neuronal activity in interconnected regions. The adaptation of eye movements when reversing spectacles are worn is explained by highly predictable modifications in very specific and coordinated regions of the cerebellum and brainstem. And example can be piled upon example. (Churchland, (2002, p. 44)).

The thought is that if changes in mental states correspond in a rule governed way to changes in physical states, then it is plausible to think that the mind is nothing over and above the brain.

As Churchland shows (1986, 2002), fundamental advances in twentieth century neuroscience have produced much more evidence for this identification. At the smallest spatial increments, the biology of the neuron has been uncovered. At larger increments we now have access to technologies such as functional Magnetic Resonance Imaging that allow us to measure blood flow (and hence activity) in the brain. Finally, the pharmacological revolution has given much greater understanding into the brain's chemical processes.

In every case we have been able understand the mind better by localizing its functions. Churchland persuasively argues that, *contra* Skinner, getting inside the black box does help us to predict and control behavior.

The natural philosophy of mind suggested by these discoveries is the identification of the mind with our physical brain. Part of the job of philosophy here is to unpack exactly

what is thus affirmed, to question the extent to which neuroscience provides evidence for the identification, and to uncover the philosophical consequences. If in the last section we saw the working out of the philosophy of mind appropriate to logical positivism leading up to Daniel Dennett, here we will show how neuroscience's philosophical dialectic can be seen as leading up to the mature philosophy of Patricia Churchland.

a. Type-Type Identity Theory

Type-type identity theory, some times called "type physicalism," is the position that types of mental states can be identified with types of brain states. In our discussion we shall see that this identification can be unpacked in at least three unequivalent ways.

> Intensional Equivalence Thesis: For every mental property *M*, there exists a physical property *P* such that it is necessarily the case that, for all events *x*, *x* instantiates *M* if and only if *x* instantiates *P*.
>
> Semi-Intensional Equivalence Thesis: For every mental property *M*, it is necessarily the case that there exists a physical property *P* such that for all events *x*, *x* instantiates *M* if and only if *x* instantiates *P*.
>
> Extensional Equivalence Thesis: For every mental property *M*, there exists a physical property *P* such that for all events *x*, *x* instantiates *M* if and only if *x* instantiates *P*.

The Intensional Equivalence Thesis affirms that if we say that the mental property being-in-pain is a physical property such as the firing of a certain kind of nerve (somebody invented the name "C fiber firings" for this property, and we can go along, even though there are no such things as C fibers), then in any possible world where someone is in pain their C fibers are firing, and in any possible world where someone's C fibers are firing, they are in pain in that world. Whether or not there is any notion of property sameness over and above intensional equivalence is a very difficult metaphysical question. Surely though, if mental properties are "nothing over and above" physical properties, the Intensional Equivalence Thesis should follow. Importantly, the position prohibits non-physical creatures such as non-physical souls, as well as physical creatures with very different physical make up from us, from having the same mental properties that we have.

The Semi-Intensional Equivalence Thesis prohibits non-physical creatures such as non-physical souls from instantiating the same mental properties that we do, but does not prohibit physical creatures with a very different makeup from us from instantiating our mental properties such as the experiencing of pain (for example, think of the *Star Trek* episode where Spock mind-melds with the life form made of rock and realizes that the thing is in great pain).

The Extensional Equivalence Thesis allows possible worlds where both non-physical souls as well as physical things very different from us to experience pain.

Today, when people talk about "the identity thesis" they usually mean for it to entail the Intensional Equivalence Thesis.

As a side note, we can see how Putnam's polio discussion was itself ambiguous. Consider the following three theses.

> Intensional Equivalence Thesis': For every disease *D*, there exists a physical property *P* responsible for the symptoms of disease *D* such that it is necessarily the case that, for all people *x*, *x* has *D* if and only if *x* instantiates *P*.
>
> Semi-Intensional Equivalence Thesis': For every disease *D*, it is necessarily the case that there exists a physical property *P* responsible for the symptoms of disease *D* such that for all people *x*, *x* has *D* if and only if *x* instantiates *P*.
>
> Extensional Equivalence Thesis': For every disease *D*, there exists a physical property *P* responsible for the symptoms of disease *D* such that for all people *x*, *x* has *D* if and only if *x* instantiates *P*.

Where "shmolio" is a disease with all of the symptoms of polio, but with a different biological cause, with the first thesis we would have to say that people with shmolio do not have polio. The second is consistent with shmolio suffers having polio. The third is so weak that it doesn't prohibit possible worlds where people have polio without there being any underlying physical cause. When we study functionalism we will see that these distinctions do cut different philosophical theories at the joints.

i. Some Reasons for Type Identity Theory

In "Sensation and Brain Processes," J.J.C. gives the following argument for the identification of mental and physical proerties. He writes,

> Why do I wish [to identify sensations with brain processes]? Mainly because of Occam's razor. . . There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. That is, for a full description of what is going on in a man you would have to mention not only the physical processes in his tissues, glands, nervous system, and so forth, but also his states of consciousness: his visual, auditory, and tactual sensations, his aches and pains. That these should be correlated with brain processes does not help, for to say that they are correlated is to say that they are something "over and above." . . .So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything be explicable in terms of physics. . . except the occurrence of sensations seems to be frankly unbelievable. (Smart, (1959, pp. 169-170))

This argument is only as plausible as the view that simpler theories are more likely to be true (Occam's razor), which may in fact not apply here!

However, in addition, we might also think that Putnam's discussion in "Brains and Behavior" provides evidence for the identity theorist. Remember that Putman concluded that, just as we should think of "polio" as "whatever is responsible for such and such symptoms" we should also think of "pain" as "whatever is responsible for such and such symptoms." Identifying pain with some set of neurological properties would be one way to spell this out. Thus, the reasons Putnam gives for the falsity of logical behaviorism should be thought of as providing some evidence for Smart's position.

Third, as we saw previously, success in localizing mental activity in terms of brain function provides evidence for type physicalism. For, if type physicalism were true, then you would expect this success. Thus, the success provides inductive evidence for type physicalism.

ii. Objections and Refinements

Here we will consider four standard objections to type-type physicalism: (a) the argument from intensional inequivalence, (b) the objection from Cartesian intuitions, (c) Chalmers' zombie objection, and (d) the multiple realizability objection.

1) The Argument from Intensional Inequivalence

Smart imagines that interlocutors might put forward the following kind of refutation. Assume for *reductio* that the mental property pain is identical to some neurological process (call this process N). Then we can argue that:

(1) Jones believes that pain is generally bad.
(2) Jones does not believe that the neurological process N is generally bad.
(3) Therefore, pain cannot be identical to some neurological process.

Unfortunately, this is an exceedingly bad argument. Things can be identical at all possible worlds and *still* be such that one can have a belief about one of them and not a belief about the other. Consider the following.

(1) Jones believes that 57 is prime.
(2) Jones does not believe that 35 + 22 is prime.
(3) Therefore, 57 cannot be identical to 35 + 22.

Smart wrote his article prior to the development of rigorous theories of necessity and possibility, and he seems to assume that a sentence *P* is necessarily true if, and only if, it is not possible to rationally doubt that *P*. He writes,

> In short, the reply to Objection 1 is that there can be contingent statements of the form "*A* is identical with *B,*" and a person may well know that something is an *A*

> without knowing that it is a *B*. An illiterate peasant might well be able to talk about his sensations without knowing about his brain processes, just as he can talk about lightning though he knows nothing of electricity. (Smart, (1962, p. 171))

Thus, he seems to appeal to the weaker Extensional Equivalence Thesis, repeated here.

> Extensional Equivalence Thesis: For every mental property *M* there exists a physical property *P* such that for all events *x*, *x* instantiates *M* if and only if *x* instantiates *P*. This is consistent with possible worlds existing where there mental properties that are wholly non-physical.

As we will see, the next two objections do work against the stronger thesis and not against the weaker one. In this case however, Smart is being too quick. Our example of 57 being identical to 35 + 22 shows that there can be *necessary* statements of the form "*A* is identical with *B*," and a person may well know that something is an *A* without knowing that it is a *B*. So one can defend the stronger thesis from the kind of counterargument Smart countenances.

2) The Argument from Cartesian Intuitions

Our second objection is the one that might be attributed to Descartes.

> (1) I can conceive of existing without a body.
> (2) Conceivability implies possibility.
> (3) Therefore it is possible that I exist without a body.
> (4) But then having a body is not an essential property of myself.
> (5) If having bodies is not essential, then it is possible that there exist creatures just like ourselves but that have no bodies (angels).

Now consider, analogously, *being in pain* and the property of being a non-physical angel that can feel pain. If this is a coherent possibility, there would exist a mental property (*being in pain*) that is such that in some possible world all of the physical properties in that world will be such that there exists an event of an angel's being in pain that does not instantiate any physical properties. This second entailment is inconsistent with both the Intensional Equivalence Thesis and the Semi-Intensional Equivalence Thesis.

As an endnote, people didn't begin to take the Cartesian Intuitions seriously until the publication of Saul Kripke's classic monograph *Naming and Necessity*, where the author utilizes the theory of necessity developed in the monograph to defend an argument very similar to the above.

With the argument as we have presented it, one could either balk at the claim that we really do have a conception of a non-physical thing with a mental life, or one could argue that conceivability is no good guide to possibility. In sections 3c., 5c.iii., and 5c.iv. below these criticisms are explored.

3) Chalmers' Zombie Intuitions

In *The Conscious Mind* David Chalmers argues that:

> (1) I can conceive of a world just like ours in every physical respect, except where there do not exist any phenomenal states such as pain.
>
> (2) Conceivability implies possibility.
>
> (3) Therefore, it is possible that there exists a world just like ours in every physical respect, except where there don't exist any phenomenal states such as pain.

This is the exact dual of the Cartesian intuitions, which asserted the possibility of a non-physical thing with a mental life. Chalmers asserts the possibility of a thing physically indiscernible from us with no mental life. While the Cartesian intuitions undermine the "left to right" conditionals of the modalized theories, Chalmers seeks to undermine the right to left reading.

Then, clearly one who is moved by both the Cartesian and Chalmers type thought experiments will only assert some weaker connection between mental and physical properties, at best the Extensional Equivalence Thesis. However, as we'll see in the next section, even this is problematic.

As with the Cartesian intuitions one might argue that, *pace* Chalmers, conceivability does not entail possibility, or that we do not really have a conception of zombies the way he thinks we do.

4) Multiple Realizability

While the Cartesian argues that angels are *possible*, proponents of the multiple realizability objection can make a strong case that the phenomena of multiple realizability is an *actual* phenomena. In this manner they attempt to undermine all three forms of type physicalism.

Though the objection was initially made in Putnam's "Psychological Predicates," we can consider the formulation of the problem in the opening section of Block and Fodor 's "What Psychological States are Not." Here we go through three examples that Block and Fodor take to undermine type-type physicalism.

a) Robots

Block and Fodor consider the possibility of applying "psychological predicates" to artifacts such as robots:

> Finally, if we allow the conceptual possibility that psychological predicates could apply to artifacts, then it seems likely that physicalism will prove empirically false.

> For it seems very likely that given any psychophysical correlation which holds for an organism, it is possible to build a machine which is similar to the organism psychologically, but physiologically sufficiently different from the organism that the psychophysical correlation does not hold for the machine. (Block & Fodor, (1980, p. 238))

Again, consider *being in pain*. If it is possible that robots exist that are a part of an event of *being in pain*, then there is a mental property *M* (being in pain), such that for all physical properties *P* it is possible that there exists an event *x* (the robot being in pain), such that *x* instantiates *M* and *x* does not instantiate *P*.

b) Evolutionary Theory

Block and Fodor argue that evolutionary theory makes it more likely that psychological states could evolve with completely different underlying physiologies. They write,

> Psychological similarities across species may often reflect convergent environmental selection rather than underlying physiological similarities. For example, we have not particular reason to suppose that the physiology of pain in man must have much in common with the physiology of pain in phylogenetically remote species. But if there are organisms whose physiology is quite different, such organisms may provide counterexamples to the psychophysical correlations physicalism requires. (Block & Fodor, (1980, p. 238))

Evaluating the relevance of this would involve evaluating evolutionary explanations of, for example, the evolution of wings and eyes in different species. Interestingly, while this criticism would thus undermine all three versions of type physicalism, it might make one start to think that there is nothing particularly mystical or mysterious about the multiple realizability of mental properties *vis a vis* neurological properties. There's nothing particularly mystical or mysterious about the property *having a wing*, but if it is (by our best evolutionary story) multiply realized *vis a vis* some other set of properties, we still presumably have a roughly functionalist explanation of what a wing is and can explain for each creature with a wing how it came to have one. In section 5d.iii. we discuss further ramifications of this.

c) Brains

Block and Fodor's quick discussion of "the Lashleyan doctrine of neurological equipotentiality" is worth quoting at length.

> . . .it does seem clear that the central nervous system is highly labile and that a given type of psychological process is in fact often associated with a variety of distinct neurological structures. For example, though linguistic functions are normally represented in the left hemisphere of right handed persons, insult to the left hemisphere can lead to the establishment of these functions in the *right* hemisphere (Of course, this point is not *conclusive*, since there may be some relevant

> neurological property in common to the structures involved). (Block & Fodor, (1980, p.238))

If this is correct, and if one means by "physical property" a property that can be located by neurosurgeons (when the neurosurgeons are not allowed to query the owner of the brain in question while locating the property), then it does seem that the Lashelyan doctrine provides a lot of evidence for the falsity of the Extensional Equivalence Thesis. Consider the mental property "thinking about a refrigerator." In all likelihood, when you are thinking about a refrigerator something completely different is going on in your brain than in mine.

On this subject the neurobiologist William R. Calvin writes:

> Another reason for caution about localizing functions to particular places is that individuals are so variable. The anatomy is much more variable than textbooks tend to mention: For example, the primary visual cortex. . . varies threefold in size between seemingly normal adult humans. The motor strip isn't always in the same order, and patches of sensory cortex are sometimes found in front of motor strip, completely contrary to the motor-in-front, sensory-behind notions of the textbooks. If subdivisions as seemingly stereotyped as the "primary maps" can vary so much, we should be especially careful when generalizing about the brain's terra incognita up front. . . Neurosurgeons mapping the language cortex of individual patients have found much local specialization, though the overall map varies enormously from one patient to another.' (Calvin, (1990, pp. 63-64))

If we think of the human brain like a computer (as we shall in the next section), with its own unique hardware and software, then this is what one would expect.

Before discussing token physicalism and neurophilosophy, one must note that all of these criticisms are still somewhat controversial. For example, in *The Conscious Mind* David Chalmers argues against materialist construals of the mind without relying on the multiple realizeability considerations. In fact, in a footnote he writes,

> In the philosophical literature multiple realizability is often pointed to as the main obstacle to "reduction," but as Brooks (1994) argues, it seems largely irrelevant to the way that reductive explanations are used in the sciences. Biological phenomena such as wings can be realized in many different ways, for example, but biologists give reductive explanations all the same. Indeed, as has been pointed out by Wilson (1985) and Churchland (1986), many physical phenomena that are often taken to be paradigms of reducibility (e.g. temperature) are in fact multiply realizable. (Chalmers, (1996, p. 364))

Nonetheless, as Wilson realizes, one still wants an explanation of why, for example, temperature of a gas and temperature of a solid can both be considered instances of temperature. At least in this sense, multiple realizability places an explanatory burden on the philosophy of mind.

b. Token-Token Identity Theory

Jaegwon Kim defines a weaker position commonly referred to as "token physicalism," in this manner:

> [Token Physicalism] Every event that falls under a mental-event kind also falls under a physical event kind (or every event that has a mental property has also some physical property). (Kim, (1996, p. 59))

Thus, the token physicalist believes that every mental event token is identical to a physical event token, but for the reasons given above, refuses to identify mental event types (or mental properties) with physical event types (physical properties). It is important to realize that Token Physicalism is strictly logically weaker than Type Physicalism. So if Type Physicalism is true, then Token Physicalism is true. One can consistently deny Type Physicalism and affirm Token Physicalism.

In an inspired passage, Kim describes just how weak an affirmation one makes with token physicalism.

> . . .token physicalism is a weak doctrine that doesn't say much; essentially, it only says that mental and physical properties are instantiated by the same entities. Any event or occurrence with a mental property has some physical property or other. But the theory says nothing about the relationship between mental properties and physical properties, the relation between pains, itches, thoughts, consciousness, and the rest, on the one hand, and types of neural events on the other. Token physicalism can be true even if there is nothing remotely resembling a systematic relationship between the mental and the physical. In a world in which token physicalism is true, there can be all sorts of mental properties and characteristics, the hurting sensation of pains, the bluish gray of an afterimage, the bright red phenomenal color of a visual datum, and countless other sensory qualities, but there need be no dependencies, or even correlations, between mental and physical properties. As far as token physicalism goes, there could be another world just like it in every physical detail except that mentality and consciousness are totally absent. *Token physicalism, therefore, can be true even if mind-body supervenience fails*: What mental features a given event has is entirely unconstrained by what biological/physical properties it has, as far as token physicalism goes, and there could be a molecule-for-molecule physical duplicate of you who is wholly lacking in consciousness, that is, a zombie. This means that the theory says nothing about how mental properties of an event might be physically based or explained. Token physicalism, then, is not much of a physicalism. In fact, if we accept mind-body supervenience as defining minimal physicalism, token physicalism falls outside of the scope of physicalism altogether. (Kim, (1996, p. 61))

In this passage Kim clearly is referring to "token physicalism" as the position that refuses to affirm anything more about the connection between the mind and body than that every

event that falls under a mental-event kind also falls under a physical-event kind. Given the reasons we earlier gave for type physicalism (simplicity, localization of mental functions, and Putnam's considerations) it is clear that *mere* token physicalism is too weak a position. Materialists need to account for the mind-brain relationship in a more robust manner.

c. Neuro-Philosophy

Clearly, type-type and token-token identity theories do not exhaust the logical space of possible connections between the mind and brain. One desires a position consistent with both multiple realizability and the fact that our mental states are closely connected to our brain states. Of all twentieth century philosophers, Patricia and Paul Churchland have done the most to work out such a position. As with Dennett in the previous section, their views and developments are both too subtle and too substantive to do service to in an encyclopedia article. Here I will discuss three themes from Patricia Churchland, strategies that, if successfully prosecuted, would lead to a position satisfyingly intermediate between type-type and token-token physicalism.

First, and perhaps most controversially, Churchland argues that the fanciful possible worlds thought experiments with which we have populated our discussion thus far are irrelevant. Here current statement of this criticism is worth quoting at length. First, she argues that such possibilities are irrelevant to proper explanation.

> To most of us, this [Chalmers' zombie] argument is puzzling, because many things are logically possible but not empirically possible, such as a 2-ton mouse or a spider that can play the flute. Why should we suppose that the logical possibility of a zombie tells us anything interesting about what research could be successful? After all, what neurophilosophy is really interested in is the actual empirical world and how it works. The reply depends on the pivotal claim about the standards for an explanation, namely, that a *proper explanation must foreclose logical possibilities*.
>
> Assuming that this is pivotal claim here, we need to recognize how absurdly strong a claim it is. Not only does it rule out explaining consciousness in terms of brain function, but it also rules out explaining consciousness in terms of *soul function* or *spooky-stuff function* or *quantum gravity* or *anything else* you might think of. So strong is the demand it places on successful explanation that no scientific explanation of any phenomenon has ever met it, or ever could meet it. . . (Churchland, (2002, p. 177))

Churchland also argues that such possibilities may not be real possibilities.

> . . .*all* such "conceivability" arguments. . . want to draw an interesting conclusion about the nature of how things *really* are. *Nothing* interesting follows, however, from the fact that some particular human is, or is not, able to imagine something. That something *seems* possible does not thereby guarantee it *is* a genuine possibility in any interesting sense, so why should we think that the zombie idea *is* genuinely

> possible? To insist on its possibility on grounds that the premises are grammatical is to *confuse a real possibility with mere grammaticality*. (Churchland, (2002, p. 177))

This is fascinating precisely because it suggests a new strategy for philosophical naturalists. Traditionally, naturalists had been interested in either reducing theories to naturalistically acceptable ones (i.e. the manifest to scientific image) or refuting such theories that cannot be so reduced (e.g. the atheist's arguments against God's existence). Churchland shows that by limiting modal/possible worlds talk to that only required by natural science, the naturalist can undermine anti-naturalistic arguments. By further arguing that the anti-naturalist's possible worlds hypotheses would undermine normal science, she has shown that the restriction is warranted or at the very least put the burden of proof on the anti-naturalist. We shall return to these points in section 5c.iii. and 5c.iv.

Second, Churchland argues forcefully that much mainstream philosophy of mind has a mistaken view of scientific progress. Philosophers of mind seem to assume that the manifest image is immutable and correct, and that it is philosophy's job to explain why the manifest image is correct. By this view, the philosophy of mind should restrict itself to either showing how the manifest image can be justified by the scientific image or (barring that) should show the scientific image to be incomplete.

Churchland notes that nearly every development in science changes the manifest image substantially. For example, (Churchland, (2002, pp. 129-120)) understanding fire to be oxidation, accompanied by learning more biology and astronomy, radically changed our concept of fire. We now know that rusting, calcification, and bodily metabolism are all instances of fire, and that the sun, lightning, northern lights, fireflies, and comets are not.

The moral of this is that failure to explain some mental states in terms of physical states might be due to our false beliefs about the mental. Given the history of the progress of science, this is what one would expect and indeed what has already been observed.

> In the last fifty years, we have come to realize that epilepsy is best understood in neurobiological terms, not in terms of the divine touch. Hysterical paralysis is not dysfunction of the uterus, but of the brain. In subjects who are compulsive handwashers, possession by spirits or superego dysfunction explains and predicts far less than neuromodulator levels. The discovery that highly addictable subjects have a gene implicated in the quirks of their dopamine reward system begins to hint that we will want to reconsider what exactly having or lacking will power comes to. (Churchland, (2002, p. 32))

As with Churchland's take on thought experiments, if she is right, then the burden of proof on the materialist is lessened considerably and the philosopher's task greatened. For now the philosopher must keep abreast of relevant science and be open to critiquing the manifest image, which, one could argue, is more in line with the (pre 20th Century) traditional task of philosophy. We shall return to this train of thought in section 5e.

Finally, Churchland holds that neuroscience and cognitive psychology have much to learn from each other.

> . . .the fact is that neuroscience and cognitive science are coevolving, like it or not. This coevolution is motivated not by ideology, but by the scientific and explanatory rewards derived from the interactions. Increasingly, this trend means that data from neuroscience are having an impact on how we frame questions about the mind and how we rethink how best to characterize psychological phenomena themselves. (Churchland, (2002, p. 27))

The hope is that insofar as some mental states are multiply realized across different brains, cognitive psychology will help us identify those states and neuroscience will help us understand how the brain can embody the correct causal roles.

However, Churchland does note that the connection has been used by Fodor and others to dismiss the relevance of neuroscience to cognitive psychology and the philosophy of mind. She is adamant that our understanding of neuroscience has, and will continue to, change our manifest image and cognitive psychology itself. Fully grasping this point requires detouring through the third great twentieth century revolution. We shall discuss this point again in 5d.

4. Computer Science

The development of the modern digital computer is one of the most exciting and philosophically interesting stories in history. While influential political philosophy written by Aristotle, Locke, Voltaire, Smith, Marx, etc. can rightfully be claimed to have had tremendous influence on how modern societies organize themselves, Leibniz, Frege, Russell, Turing, Gödel, Church, and Von Neuman are perhaps the only philosophers who can claim to have birthed a world changing technology.

Earlier, while discussing Logical Positivism, we noted how the new logic developed by Russell and Frege played a key role in the explanation of knowledge of both empirical and analytic truths. In the former case, logic is used to derive testable predictions from physical theories and in the latter to derive theorems from mathematical axioms.

In that discussion we had no need to broach the reason why anyone would want to characterize mathematics in such a manner. However, since it was the attempt to be clear about this motivation that gave rise to the digital computer, we do well to discuss it here.

If one's logic had the right properties, then the certainty of a finitely specifiable set of meaning characterizing axioms could be transmitted to a certainty about an infinite set of theorems following logically from those axioms. Since a mathematical theory just is a set of theorems, the logic and axioms would provide the foundation for the theory.

Now, which properties must logical proofs have for the logic to play this foundational role? The great hope (at the beginning of the twentieth century) was for each step in a

proof from axioms to a theorem to be clearly and indubitably valid. A system of deductive logic then is a system of such possible steps. For this to be successful it must be the case that a purported proof is mechanically checkable, or computable. For if a machine could in principle check each step of a purported proof for validity, then we can be absolutely certain that the proof is correct (*mod* the correctness of the deductive principles of the logic).

A complete discussion of the historic, philosophical, and mathematical context of this foundational quest would take us beyond our current context (see (Shapiro, 2000) for an excellent account). What is important for our purposes is that early twentieth century logicians, philosophers, and mathematicians such as Hilbert, Russell, Frege, and Gentzen did develop logics that are such that the purported proofs in these logics can be checked.

As this was being achieved, one started to think of how much math could be given such an axiomatic status. It is in this context that Gödel's 1920s results shook the world. Gödel proved that any attempted axiomatization of number theory would be (if consistent) incomplete, in the sense that true sentences of number theory would not be provable from the purported axioms. This result is world shaking as it caused obvious severe (perhaps fatal) problems for the foundational enterprise (see section 5d.i. for a discussion of other possible implications).

However, equally world shaking are the techniques Gödel developed in coming up with his proof. In thinking through his result, Gödel had to make mathematically precise the property of being axiomatizable, that is, the property of being such that proofs in a system are mechanically checkable. This required precisely defining the set of primitive recursive, and then later recursive, functions. Then a set of numbers is mechanically checkable if, and only if, the characteristic function for the set (one that delivers "1" for a member of that set, and "0" for a non-member of that set) is recursive. When this is added to Gödel's technique (now called Gödel numbering) of assigning unique prime numbers to sentences in a logical language, then we can precisely define a set of sentences as being mechanically checkable/computable, if, and only if, its set of Gödel numbers are.

Other thinkers besides Gödel came up with quite distinct ways to characterize the computability of a set of numbers. Of these, the most important were Church's lambda definability (with the lambda calculus ultimately making a tremendous impact on linguistic semantics via Richard Montague's work), and Turing's notion of machine computability (to be described in the next section). Amazingly, all of these on the surface quite different notions ended up being equivalent. One can prove that a set of numbers is recursive if, and only if, it is lambda definable, if and only if, it is Turing machine computable. These results led Alonzo Church to theorize that he, Gödel, and Turing must have correctly analyzed the notion of computability. If in such different ways they ended up surprisingly characterizing the same sets of numbers, then it is plausible to think that each was correct. This is now known as Church's Thesis. Given that Turing machine computability is equivalent to the Von Neuman computability of contemporary computers, Church's Thesis is another way of saying that if a function is

intuitively computable at all, then it can be computed by a modern digital computer (given enough space, time, and energy resources).

In addition to designing the first computing machine, Turing (1950) devised a fascinating test for whether a machine could correctly be said to possess intelligence. A simplified version of this involves placing a person with a keyboard and a monitor on one side of a screen. Alternatively let her interact with messages posted by the computing machine and messages posted by a human. If, after many sessions, the person cannot guess better than chance whether the interlocutor is the person or the computing machine, the computing machine can be said to be intelligent.

This manner of operationalizing intelligence immediately suggests a theory of mind. If an intelligent (by Turing's test) machine really is thinking, then perhaps our thinking is nothing over and above computation? Moreover, by Church's Thesis, such computation would be in principle no different from what computing machinery can already do. So perhaps the brain is nothing but a biological computing machine and thinking analogous to the running of software on that machine. This thought then leads to two directions: (1) the attempt to fully develop artificial intelligence, the building of thinking machines, and (2) the attempt to fully develop a theory of mind along the lines suggested by the metaphor. Immediately, one sees the advantage over type-type and token-token identity theories. Unlike type physicalism, one would expect multiple realizability, as the same software can run on different hardware. Unlike token physicalism, the theory is constrained in the sense that the physical causal structure must be able to perform the computations in question.

For a fascinating and accessible account of the philosophical, logical, mathematical history discussed above, see (Davis, 2000). For the canonical presentation of the logical results cited above, see (Boolos, Burgess, & Jeffrey, 2002).

a. Functional State Identity Theory (Varieties of Turing Machine Functionalism)

Most of this section will be devoted to Hilary Putnam's functional state identity theory, as it was the earliest version of the computational theory of mind, and hence the recipient of most of the canonical criticisms of that theory. It is also the one tied in the strongest manner to Turing machines.

i. Standard Turing Machines

In our treatment, a standard Turing machine consists of a piece of tape infinitely long in both directions (though we could restrict it to being infinitely long in merely one direction without losing anything of import) which is divided into an infinite number of cells, each with either a 0 or a 1 written on the cell, and a counter which follows a given set of instructions, in our case to either write a 0 or a 1 on the tape or to move left or right. For example, if we denote the location of the counter by boldfacing the number in

a tape's cell, the following could illustrate the execution of a Turing machine program.

. . . . .	0	0	0	**0**	0	0	. . . . .
. . . . .	0	0	0	**1**	0	0	. . . . .
. . . . .	0	0	**0**	1	0	0	. . . . .
. . . . .	0	0	**1**	1	0	0	. . . . .
. . . . .	0	**0**	1	1	0	0	. . . . .
. . . . .	0	**1**	1	1	0	0	. . . . .

Each row represents the tape. For the second to sixth row, the tape has just been affected either by the counter moving to a new place, or by the counter writing a new number (nothing prohibits the counter from writing the same number over the old one, in this case two rows would look exactly alike).

In our treatment, a standard Turing machine program will consist in a series of commands, all of the following form.

If in state *n* and the counter reads *m*, then write/move *o* and then go to state *p*.

Here the variables *n*, and *p* range over natural numbers, *m* can be equal to either 1 or 0, and *o* can be equal to either 1, 0, L (for left), or R (for right). The only restriction we impose is that each line be well defined; we prohibit multiple lines that give contradictory instructions. Thus, the following is *not* a Turing machine program.

If in state 1 and the counter reads 1, then write 0 and then go to state 1.
If in state 1 and the counter reads 1, then move L and then go to state 2.

The machine can receive at most one instruction for where to move or what to write given whatever the counter is reading in a given state. Likewise, the machine can receive at most one instruction for what state to move into given the state it is in, what the counter has read, and what it has written or moved. So the following is also *not* a Turing machine program.

If in state 1 and the counter reads 1, then write 0 and then go to state 1.
If in state 1 and the counter reads 1, then write 0 and then go to state 2.

Here is an acceptable Turing machine program, a program that makes the machine do what is executed by the above portrayal of a series of tapes.

If in state	and the counter reads	then write/move	and then go to state
1	0	1	1

1	1	L	2
2	0	1	2
2	1	L	3
3	0	1	3

program 1: standard Turing machine that prints three ones to the left of the initial counter place

We can show how this program executes *via* another table. In the following, the center 8 columns of the table represent eight places on the tape (remember that there are really an infinite number of cells on the tape in both directions), and a number is boldfaced if the counter is in the cell in which that number occurs. The first row shows the initial state of the tape, and the program begins to execute with the command "go to state 1." Then the cell on the rightmost column tells the machine how to alter the tape, given the commands of the state the machine is in as well as the number written in the cell the counter was previously in.

	0	0	0	0	0	0	0	0	go to state 1
since the counter reads 0, write 1	0	0	0	0	1	0	0	0	go to state 1
since the counter reads 1, move L	0	0	0	0	1	0	0	0	go to state 2
since the counter reads 0, write 1	0	0	0	1	1	0	0	0	go to state 2
since the counter reads 1, move L	0	0	0	1	1	0	0	0	go to state 3
since the counter reads 0, write 1	0	0	1	1	1	0	0	0	

execution of program 1 when beginning with blank tape

Turing machine programs can be represented in a number of other ways. For example, consider the following.

	0	1
1	1:1	L:2
2	1:2	L:3
3	1:3	

machine table for program 1

ii. Probabilistic Turing Machines

In "The Nature of Mental States" Hilary Putnam first proposed thinking of psychological systems like humans as very complicated Turing machines. To understand what Putnam had in mind we must generalize Standard Turing Machines. First we generalize the standard Turing machine to probabilistic Turing machines. For simplicity's sake we assume that the tape of this machine has a beginning and is infinite in only one direction,

towards the right. Then a probabilistic Turing machine program is a set of commands of the following sort.

If in state n and the counter reads m,
then o_1 percent of the time write/move p,
and then q_1 percent of the time go to state r_1,
:
q_n percent of the time go to state r_n.
:
then o_n percent of the time write/move p_n
and then s_1 percent of the time go to state t_1,
:
s_n percent of the time go to state t_n.

We have a couple of restrictions in force here.

First, for any o_i, p, q_1, . . .q_n, and r_1, . . .r_n, such that we have a line containing

then o_i percent of the time write/move p,
and then q_1 percent of the time go to state r_1,
:
q_n percent of the time go to state r_n,

then none of the r_is are equal to one another, and q_1+. . . + qn = 100%.

Second, for any n, m, o_1, . . .o_n, q_1, . . .q_n, and r_1, . . .r_n, when there are a numbered list of commands of the form,

if in state n and the counter reads m,
then o_1 percent of the time write/move p_1,
and then q_1 percent of the time go to state r_1,
:
q_n percent of the time go to state r_n,
:
then o_n percent of the time write/move p_n
and then s_1 percent of the time go to state t_1,
:
s_n percent of the time go to state t_n,

then none of the p_is are equal to one another, and o_1+. . . + o_n = 100%.

To best see how these restrictions work we can write some programs for Probabilistic Turing Machines.

First, to see how this notion is a proper generalization of the standard Turing machines we can rewrite program 1 in terms of program for a probabilistic Turing machine (here

we go back to letting the tape go to the left and the right of the starting state).

If in state	and the counter reads	then	percent of the time write/move	and then	percent of the time go to state
1	0	100	1	100	1
1	1	100	L	100	2
2	0	100	1	100	2
2	1	100	L	100	3
3	0	100	1	100	3

program 1: probabilistic Turing machine that prints three ones to the left of the initial counter place

Then to see how our restrictions are in force, consider the following program.

If in state	and the counter reads	then	percent of the time write/move	and then	percent of the time go to state
1	0	25	R	30	1
				70	1
		75	1	100	2
1	1	100	0	40	2
				60	3
2	0	100	1	100	2
2	1	50	R	100	3
		50	1	25	1
				25	3
				50	1
3	1	100	0	100	3

iii. Probabilistic Automata

Putnam's notion of a probabilistic automaton is arrived at by generalizing the notion of a probabilistic Turing machine. Again a probabilistic Turing machine program is a set of commands of the following sort:

If in state n and the counter reads m,
then o_1 percent of the time write/move p,
and then q_1 percent of the time go to state r_1,
:
q_n percent of the time go to state r_n.
:
then o_n percent of the time write/move p_n

and then s_1 percent of the time go to state t_1,
:
s_n percent of the time go to state t_n.

where the same restrictions as before are in force. The only differences are that the *m* variable is thought to stand not for 0s and 1s but rather for sensory inputs, so that instead of the input variable "the counter reads *m*" we get "the organism's sensory inputs are in state *m*", and likewise the "write/move *p*" command becomes "produce motor output *p*." Likewise instead of the program running on a tape, this program runs on a psychological organism's body.

This is just the result of generalizing standard Turing machines twice: first to allow probabilities in, and second to think of the machine itself as being a biological organism, whose sets of inputs and outputs are sensory inputs and motor outputs respectively. With this understanding, we can understand Putnam's claim about pain. Interestingly, as long as there are a finitely specifiable set of sensory inputs and motor outputs, Putnam's probabilistic automata can be modeled by a digital computer, and hence by a standard Turing machine.

The hypothesis that "being in pain is a functional state of the organism" may now be spelled out more exactly as follows:

> (a) All organisms capable of feeling pain are Probabilistic Automata.
>
> (b) Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e. being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).
>
> (c) No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in (b).
>
> (d) For every Description of the kind referred to in (b), there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset. (Putnam, (1980b, p. 226))

iv. Objections to Functional State Identity Theory

In "What Psychological States are Not," Block and Fodor present six arguments against the theory that a type of mental state can be identified with a type of probabilistic automata state. Here we consider four of them.

Block and Fodor do not concern themselves with any deep metaphysical issues concerning the identity, but rather criticize a biconditional which should follow from the identity:

> *O* is in such and such a type of psychological state at time *t* if and only if *O* is in such and such a type of machine table state at time *t*. (Block and Fodor, (1972, p. 240))

They are not criticizing the mere claim that one can describe the behavior of a psychological agent *via* probabilistic automata.

> There are thus two important respects in which FSIT involves more than the claim that organisms which satisfy psychological predicates have descriptions. First, FSIT claims that such systems have unique best descriptions. Second, FSIT claims that the types of machine table states specified by the unique best description of a system are in correspondence with the types of psychological states that the system can be in. (Block and Fodor, (1980, p. 241))

Remember that the functional state identity theory is supposed to be better than behaviorism because the inputs to the state of a probabilistic automaton include not only sensory inputs, but other psychological states as well. The theory is supposed to be better than type identity theory because it is supposed to explain how the same mental state can be realized by different physical states. As Block and Fodor show, however, the theory has a whole host of other problems.

1) Occurrent Versus Dispositional Psychological Properties

Block and Fodor claim that functional state identity theory cannot account for the distinction between occurrent and dispositional psychological properties. Their argument can be presented in the following manner:

> (1) By the functional state identity theory, a person is in an occurrent mental state ("sensations, feelings, thoughts, and so on") like being-in-pain if the best description of her behavior is one where a probabilistic automaton is executing the commands given in the probabilistic automaton state corresponding to pain.
>
> (2) Presumably, a person possesses a dispositional state ("beliefs, desires, inclinations, and so on") such as liking-strawberry-ice cream if the best description of her psychology is one with a state corresponding to liking-strawberry-ice-cream.
>
> (3) But then for each disposition, there will be a corresponding occurrent mental state that occurs when the best description of her behavior is one where a probabilistic automaton is executing the commands given in the probabilistic automaton state corresponding to that disposition.
>
> (4) But there is likely no intersubjectively or intrasubjectively unique occurrent mental state for the vast majority of dispositional properties. Think of the vast majority of the ways one can manifest one's disposition to believe a proposition *P*.
>
> (5) Thus, the Functional State Identity Theorist can either deny that dispositional psychological properties are genuine psychological types, or maintain that there

> really is a unique occurrent mental state that occurs when each dispositional property is manifested.

Thus, the theory would need resources for characterizing dispositional states in terms of propensities to go into the relevant occurrent states and behaviors.

It's not clear that Block and Fodor aren't being a bit unfair here. They do note that this is difficult for any theory. Thus, the defender of functional state identity theory might argue that it is not a special difficulty for her.

2) Interactions Between Psychological States

Block and Fodor argue that, contrary to initial appearances Functional State Identity Theory cannot account for the fact that behavior is characteristically the product of interactions between psychological states.

> Because Functional State Identity Theory accounts for the way in which behavior can be a product of a series of psychological states, it seems to be much better suited than behaviorism.
>
> However, behavior is often the result of interactions between simultaneous mental states.
>
> Probabilistic Automata only compute things in a serial fashion. Therefore, Functional State Identity Theory cannot account for how our behavior is often in virtue of two or more simultaneous mental states.
>
> Indeed FSIT even fails to account for the fact that an organism can be in more than one occurrent psychological state at a time, since a probabilistic automaton can be in only one machine table state at a time. (ibid., p. 243)

Thus, to the extent that the computational metaphor is a useful one, we want to characterize the mind as divided into "modules," whose computational processes run in parallel with one another, but also interact in various ways. Fodor's own *The Modularity of Mind* argues forcefully for this.

Thus, the theory would at best need to be generalized to such that people are characterized as a set of Probabilistic Automata operating at the same time.

3) The Individuation of Psychological Types

Block and Fodor argue that the Functional State Identity Theory incorrectly individuates psychological types.

> Two machine table states of probabilistic automata differ if they differ in their range of outputs, or in their range of successor states, or in the probability distributions associated with either of these ranges. (ibid., pp. 245-246))

But,

> if we transfer this convention for distinguishing machine table states to the type identification of psychological states, we get identity conditions which are, as it were, too fine-grained. (ibid., p. 246))

That is, mental states that are type identical end up being treated as distinct by the psychological theory.

For example, the fact that you are prone to say "darn" when stubbing your toe, and the fact that I never say "darn" makes it the case that we are not both in the same type of mental state when we stub our toes. We want to say that we are both in pain, but functional state identity theory precludes this.

Block and Fodor point out that this problem generalizes. Indeed, on the assumption that there is a computational path from every state to every other, any two automata that have less than all their states in common will have none of their states in common.

Perhaps the defender of the Functional State Identity Theory could appeal to a notion of similarity of machine table state to get around this. Thus, instead of,

> *O* is in such and such a type of psychological state at time *t* if and only if *O* is in such and such a type of machine table state at time *t*. (ibid., p. 240))

we would get

> *O* is in such and such a type of psychological state at time *t* if and only if *O* is similar in the right respects to such and such a type of machine table state at time *t*.

Block and Fodor seem aware of this, when they cite Putnam's claim that "the difficulty of course will be to pass from models of specific organisms to a normal form for the psychological description of organisms." Machine table states for specific organisms would have to be relevantly similar to the normal form for them to possess mental properties. This, however, is a bugbear. Since, anything is in some respects similar to anything else, the functional state identity theorist now has a very difficult story to tell.

4) Productivity

Block and Fodor argue that the Functional State Identity Theory cannot account for productivity.

> If we abstract away from issues concerning memory limitations, then a person has an infinite number of mental states arrived at of the form "the belief (thought, desire, hope, and so forth) that S."

But this is inconsistent with the claim that possessing a given psychological property, such as believing that 1 + 1 = 2, is identical with having a certain machine table state in the best description of one's psychology in terms of probabilistic automata. This is because any such automaton is finite.

To point out that we only have a finite memory is to no avail, as

> . . .the fact that there are arithmetical statements that it is nomologically impossible for any person to believe is a consequence of the character of people's memory, not a consequence of their mental representation of arithmetic. (ibid., p. 241))

The authors reiterated that they are here just reducing to absurdity the claim that mental types are identical to machine table states. They are not denying that something computational goes on in your mind when you add numbers together, nor that this can be described using computational metaphors.

v. Tentative Conclusions

So we can conclude that using Turing machines to model mentality requires considerably more subtlety than the mere identification of types of mental states with states of a probabilistic automaton. As a consequence of Fodor and Lepore's criticism, the best account one can come up with is a vague promissory note involving a "normal form" description of modularized probabilistic automata. While the original functional state identity theorist gives us

> *O* is in such and such a type of psychological state at time *t* if and only if *O* is in such and such a type of machine table state at time *t*. (ibid., p. 240))

at best we can now get

> *O* is in such and such a type of psychological state at time *t* if and only if *O* is similar in the right respects to such and such a type of modularized probabilistic automaton executing such and such steps at time *t*.

Of course this is incredibly vague. However, many strands of cognitive science might be thought of as fulfilling this. For example (Thagard, 1996) discusses computational models of language, reason, concepts, analogy, and vision. All of the models put forward might be thought of as providing evidence for instances of the above schema. Thus, minimally, the computational theory continues to have strong applicability in cognitive science and artificial intelligence.

b. Ramsey Sentence Functionalism

Given the upshot of the discussion in section 4a., one way to think about the Ramsey-Lewis Method is as a proposal concerning constraints on the computational architecture of a normal form modular probabilistic automaton corresponding to a given mental state. Strangely, the Ramsey-Lewis method of doing this might be taken to verify analytic behaviorism's attempt to characterize mental states merely in terms of environmental inputs and behavioral outputs. Unfortunately though, Jaegwon Kim argues forcefully (and in manner analogous to Putnam's earlier arguments about the individuation of psychological types) that this kind of functionalism doesn't work. Prior to defending this claim, we must show how the Ramsification works.

Ramsification is a simple two-step process to define some of the predicates of a first order theory in terms of the other predicates.

Step one involves taking the first order theory and existentially generalizing all of the predicates you want to define. To illustrate this we will use Kim's examples. So let the first order theory be the following

> (T) for any x, ((if x suffers tissue damage and is normally alert), x is in pain); (if x is awake, x tends to be normally alert); (if x is in pain, x winces and groans and goes into a state of distress); and (if x is not normally alert or is in distress, x tends to make more typing errors).
> (Kim, (1996, p. 105), parenthesis added)

This sentence can easily be formulated in something like first order logic. Where "$\forall$" means "for all," "$\exists$" means "there exists," "$P \rightarrow Q$" means "if P then Q," "$\wedge$" means "and," "$\vee$" means "or," "$\neg$" means "it is not the case that," we can translate the above into,

$$\begin{aligned}(T)\quad \forall x[(&(\text{ suffers tissue damage}(x) \wedge \text{normally alert}(x)) \rightarrow \text{in pain}(x))\\ &\wedge\ (\text{awake}(x) \rightarrow \text{tends to be normally alert}(x))\\ &\wedge\ (\text{in pain}(x) \rightarrow ((\text{winces}(x) \wedge \text{groans}(x))\\ &\qquad\qquad \wedge\ \text{goes into a state of distress}(x))\\ &\wedge ((\neg\text{normally alert}(x)\ \vee\ \text{in distress}(x))\\ &\qquad\qquad \rightarrow \text{tends to make more typing errors}(x))].\end{aligned}$$

Then our first step involves existentially generalizing out. The only thing that is important here is to pick different variables for each different mental predicate. From Kim again we have.

> (TR) There exists states M_1, M_2 and M_3 such that for any x, ((if x *suffers tissue damage* and is in M_1), x is in M_2); (if x *is awake*, x tends to be in M_1); (if x is in M_2, x *winces* and *groans* and goes into M_3); and (if x is not in M_1 *or* is in M_3, x *tends to make more typing errors*).
> (Kim, (1996, p. 105), parenthesis added)

In our logical language this gives us.

(TR) $\exists M_1 \exists M_2 \exists M_3 [\ \forall x [$
$((\text{ suffers tissue damage}(x) \wedge M_1(x)) \rightarrow M_2(x))$
$\wedge\ (\text{awake}(x) \rightarrow \text{tends to be } M_1(x))$
$\wedge\ (M_2(x) \rightarrow ((\text{winces}(x) \wedge \text{groans}(x)) \wedge \text{goes into } M_3(x))$
$\wedge ((\neg M_1(x) \vee M_3(x))$
$\rightarrow \text{tends to make more typing errors}(x))\ ]\ ]$

Following Kim, we'll abbreviate this with $\exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(M_1, M_2, M_3)]$. Then for the second step we define all of the predicates we have existentially quantified over in the following manner.

y is in pain $=_{\text{def}}\ \exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(\mathrm{M}_1, M_2, M_3) \wedge M_2(y)]$
y is in normally alert $=_{\text{def}}\ \exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(\mathrm{M}_1, M_2, M_3) \wedge M_1(y)]$
y is in distress $=_{\text{def}}\ \exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(\mathrm{M}_1, M_2, M_3) \wedge M_3(y)]$

To make extra sure that we are clear about this we will unpack the definition of being in pain according to our theory. First, given our abbreviation our definition is equal to

y is in pain $=_{\text{def}}\ \exists M_1 \exists M_2 \exists M_3 [\ \forall x [$
$((\text{ suffers tissue damage}(x) \wedge M_1(x)) \rightarrow M_2(x))$
$\wedge\ (\text{awake}(x) \rightarrow \text{tends to be } M_1(x))$
$\wedge\ (M_2(x) \rightarrow ((\text{winces}(x) \wedge \text{groans}(x)) \wedge \text{goes into } M_3(x))$
$\wedge ((\neg M_1(x) \vee M_3(x))$
$\rightarrow \text{tends to make more typing errors}(x))\ ] \wedge M_2(y)]$

Then, in the less formal jargon of Kim's example, we get,

y is in pain $=_{\text{def}}$ There exists states M_1, M_2 and M_3 such that for any x, ((if x *suffers tissue damage* and is in M_1), x is in M_2); (if x *is awake*, x tends to be in M_1); (if x is in *M2,x winces* and *groans* and goes into M_3); and (if x is not in M_1 *or is* in M_3, x *tends to make more typing errors*); and y is in M_2]

In this manner Ramsification provides a natural way to cash out the functionalist's claim that a psychological state can be identified in terms of its inputs and outputs, where both inputs and outputs are allowed to be physical and mental, and the *if. . . then* clauses are understood to be transitions from inputs to outputs. Ramsey proved that for a given first order theory like (T), (TR) is logically equivalent to (T) as far as entailments not involving the vocabulary quantified out. In other words, if you could do this with all mental predicates simultaneously, the resulting theory would play exactly the same inferential role *vis a vis* physical predictions.

i. A Reason for Ramsey Sentence Functionalism

Now we can make sense of the thought that the Ramsey-Lewis Method is a proposal concerning some of the constraints on the computational architecture of a normal form modular probabilistic automaton corresponding to a given mental state. Given that the *if. . . then* clauses can be thought of as transitions from inputs to outputs, the theory codified by the Ramsey Sentence can be thought of as very directly constraining our normal form automaton as one that describes the input/output conditions given by the theory. While this is still grossly speculative, it is less vague than our earlier view. We can give the new view as

> *O* is in psychological state *P* at time *t* if and only if *O* is similar in the right respects to a modularized probabilistic automaton which correctly models TR (the Ramsey sentence for the correct psychological theory of creatures of *O*'s sort) executing the steps corresponding to the state which defines *P* (using TR as is normal in the Ramsey-Lewis method) at time *t*.

In this manner, Ramsey sentence functionalism is not so much a competitor to computationalist functionalism, but rather a way to specify the computational connections at an appropriate level of generality.

ii. Kim's Objection from the Individuation of Psychological Types

One should immediately begin to wonder exactly what gets Ramsified. Following Lewis, we might think that the shared platitudes we all accept concerning mentality form a theory ripe for Ramsification. On the other hand, one might think that the theory to be Ramsified is the theory arrived at by psychologists.

One must realize that if any one of the sentences in our initial theory T is false, then there is no guarantee that (TR) ends up being true. But if (TR) ends up being false, then all of the proposed definitions fail. To see why this is the case, consider the shorthand version of Kim's pain theory,

$$\exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(\mathrm{M}_1, M_2, M_3)]$$

If one of the original statements in T is false then there probably won't exist properties M_1, M_2, and M_3 such that $[\mathrm{T}(M_1, M_2, M_3)]$ is true. But then if $[\mathrm{T}(M_1, M_2, M_3)]$ is false, our definitions for mental properties involving this sentence will always have as their extension the empty set. If $\exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(M_1, M_2, M_3)]$ is false, it follows that no y is such that $\exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(M_1, M_2, M_3) \,\&\, M_2(y)]$. But then our definition for pain, y is in pain $=_{\mathrm{def}} \exists M_1 \exists M_2 \exists M_3 [\mathrm{T}(M_1, M_2, M_3) \,\&\, M_2(y)]$, would entail that nothing is in pain. Likewise for our other definitions. From this, Kim concludes,

> This means that we had better make sure that the underlying theory is true. If our *T* is to yield our psychological concepts all at once, it is going to be a long conjunction of myriad psychological generalizations, and even a single false component will render the whole conjunction false. So we must fact this question: What is going to

> be included in our *T*, and how certain can we be that *T* is true? (Kim, (1996, pp. 108--109))

Now suppose that two psychologists dispute about some psychological truth. Potentially there is big trouble here.

> These reflections lead to the flowing thought: On the Ramsey-Lewis method of defining psychological concepts, every dispute about underlying theory *T* is going to turn out to be a dispute about psychological concepts. This creates a seemingly paradoxical situation: If two psychologists should disagree about some psychological generalization that is part of theory *T*, which we could expect to be a very common occurrence, this would mean that they are using different sets of psychological concepts. But this would seem to imply that they could not really disagree, since the possibility of disagreement presupposes the sharing of the same concepts. How could I accept and you reject a given proposition unless we shared the concepts in terms of which the proposition is formulated?
> (Kim, (1996, p. 109))

This is a big problem. The Ramsey Sentence functionalist cannot appeal to academic psychologists for the theory to be Ramsified. On the other hand, if we use commonsense folk psychology, then it isn't clear how we can believe scientific psychology to ever show folk psychology to be flawed. But in fact this does happen. However, we again get the problem that once one part of the Ramsified theory is shown to be false, the whole theory is false.

c. Causal Role Functionalism

In "Mad Pain and Martian Pain" David Lewis describes two desiderata for any philosophical account of mental states such as pain. A theory of mental states must not prohibit the existence of: (1) madmen who feel pain, but have different stimulatory inputs and motor outputs as well as different functional role with respect to other mental states, and (2) Martians who again feels pain, but whose physical realization of pain is greatly different from ours but has similar functional role.

Later in the article Lewis talks about mad Martians whose pain has a different causal role from ours (different stimulatory inputs and motor outputs) and whose physical realization of pain is greatly different from ours. Lewis writes,

> The lesson of mad pain is that pain is associated only contingently with its causal role, while the lesson of martian pain is that pain is connected only contingently with its physical realization. How can we characterize pain *a priori* in terms of causal role and physical realization, and yet respect both kinds of contingency? (Lewis, (1980, p. 216))

Since type identity theory gets mad pain correct, but goofs up on the Martian pain, and simple behaviorism and functionalism goof up on Martian pain, Lewis states:

> It seems that a theory that can pass our test will have to be a mixed theory. It will have to be able to tell us that the madman and the Martian are both in pain, but for different reasons: the madman because he is in the right physical state, the Martian because he is in a state rightly situated in the causal network. (ibid., p. 216)

Lewis considers two ways to do this. The first would be to simply disjoin the identity theory with functionalism. Thus, where one's identity theory identified pain with physical state *P*, and one's functionalism identified pain with functional role *R*, one could simply say that someone is in pain if and only if they either are in physical state *P* or in functional role *R*. [Note: our mad Martian would be a counterexample to this strategy.] The other would be to say that the word "pain" is ambiguous between its functional role meaning and its physical state meaning.

Lewis holds that a variant of the ambiguity postulation strategy can work well.

> As you'll see, I shall defend a version of it [the ambiguity thesis]. But it's not plausible to cook up an ambiguity *ad hoc* to account for the compossibility of mad pain and Martian pain. It would be better to find a widespread sort of ambiguity, a sort we would believe in no matter what we thought about pain, and show that it will solve our problem. This is my plan. (ibid., p. 216)

Lewis, in common with the early Putnam, holds that a psychological state such as pain is best characterized as a state that occupies a certain causal role, where the causes and effects of a creature's being in this state can be both mental and physical states. Since the end result of Lewis' view of this position is surprising, he is worth again quoting at length.

> If the concept of pain is the concept of a state that occupies a certain causal role, then whatever state does occupy that role is pain. If the state of having neurons hooked up in a certain way and firing in a certain pattern is the state properly apt for causing and being caused, as we materialists think, then that neural state is pain. But the concept of pain is not the concept of that neural state ("The concept of . . ." is an intensional functor.) The concept of pain, unlike the concept of that neural state which in fact is pain, would have applied to some different state if the relevant causal relations had been different. Pain might not have been pain. The occupant of the role might have not occupied it. Some other state might have occupied it instead. Something that is not pain might have been pain. (ibid., p. 216))

To understand Lewis, we must take this modal talk about the occupant of the role not necessarily occupying it very seriously. In fact, we can do this by recapitulating the distinctions we originally made when disambiguating Putnam's discussion of diseases in various ways.

Lewis' general claim can be given in this manner.

For all psychological properties *M*, there exists a set of causal inputs and outputs *C*, such that it is necessarily the case that for all events *x*, *x* instantiates *M* if and only if there exists a physical property *P* such that *P* is a *realization base* for *C* and *x* instantiates *P*.

Where "$\forall$" means "for all," "**[ ]**" means "it is necessarily the case that," "$\exists$" means "there exists," "$\leftrightarrow$" means "if and only if," "$\wedge$" means "and," "*M*" denotes an arbitrary mental property, "*C*" denotes an arbitrary causal role (set of inputs and outputs), and "*P*" denotes an arbitrary physical property, we can restate this as

$\forall M \ \exists C$ [ ] $\forall x$ ($Mx \leftrightarrow \exists P$ (realization base(P,C) $\wedge$ $P(x)$))

Now if we instantiate this universal generalization with the property of being in pain we get,

There exists a set of causal inputs and outputs *C*, such that it is necessarily the case that for all events *x*, *x* instantiates *pain* if and only if there exists a physical property *P* such that *P* is a *realization base* for *C* and *x* instantiates *P*.

$\exists C$ [ ] $\forall x$ (pain(x) $\leftrightarrow \exists P$ (realization base(P,C) $\wedge$ $P(x)$))

Now let's pick an arbitrary name for the set of inputs and outputs that make up the causal role for pain (call this *set 57*) so we can get,

It is necessarily the case that for all events *x* instantiates *pain* if and only if there exists a physical property *P* such that *P* is a *realization base* for *set 57* and *x* instantiates *P*

[ ] $\forall x$ (pain(x) $\leftrightarrow \exists P$ (realization base(P, set 57) $\wedge$ $P(x)$)

Finally, let's let the event in question be *Fred the Martian at time t*:

It is necessarily the case that *Fred the Martian at time t* instantiates *pain* if, and only, if there exists a physical property *P* such that *P* is a *realization base* for *set 57* and *Fred the Martian at time t* instantiates *P*.

[ ] (pain(Fred the Martian at time *t*) $\leftrightarrow \exists P$ (realization base(P, set 57) $\wedge$ P(Fred the Martian at time *t*))

Let's also consider what we get if we plug in *Billy the Human at time t*:

It is necessarily the case that *Billy the Human at time t* instantiates *pain* if and only if there exists a physical property *P* such that *P* is a *realization base* for *set 57* and *Billy the Human at time t* instantiates *P*.

[ ] (pain(Billy the Human at time *t*) $\leftrightarrow \exists P$ (realization base(P, set 57)

$\wedge$ P(Billy the Human at time t))

Thus, the statement captures Lewis' modal intuitions about pain; nothing prohibits Billy's realization base from differing from Fred's, or for that matter, Billy's realization base at one time from differing from his realization base at another time.

Generalized, the position we attribute to Lewis can be given in this manner:

> For all psychological properties *M*, there exists a set of causal inputs and outputs *C*, such that for all events *x*, *x* instantiates *M* if and only if there exists a physical property *P* such that *P* is a *realization base* for *C* and *x* instantiates *P*.

$\forall M \exists C [\ ] \forall x (M(x) \leftrightarrow \exists P(\text{realization base}(P,C) \wedge P(x)))$

In the case of "pain" this gets us,

> There exists a set of causal inputs and outputs *C*, such that it is necessarily the case that for all events *x*, *x* instantiates *pain* if and only if there exists a physical property *P* such that *P* is a *realization base* for *C* and *x* instantiates *P*.

$\exists C [\ ] \forall x (\text{pain}(x) \leftrightarrow \exists P(\text{realization base}(P,C) \wedge P(x)))$

This seems to capture exactly both Lewis' view that the concept pain is necessarily identical with the state occupying a causal role, as well as the view that,

> the concept and name of pain contingently apply to some neural state at this world, but do not apply to it at another. (Lewis, (1978, p. 218))

It should also be clear that this formula is consistent with the Martian pain part of the *Gedankenexperiment*. Lewis writes,

> Human pain is the state that occupies the role of pain for humans. Martian pain is the state that occupies the same role for Martians. (ibid., p. 218))

However, one might reasonably ask if our formula is consistent with mad pain. On the face of things it is not. The mad person is in pain, but is in a state that occupies a very different causal role. That is, the physical state of the mad person while the mad person is in pain, is a physical realizer of a very different causal role. So now it seems absolutely false to say what Lewis is saying. Lewis gets out of this problem in this manner:

> The thing to say about mad pain is that the madman is in pain because he is in the state that occupies the causal role of pain for the population comprising all mankind. . . We may say that *X* is in pain *simpliciter* if and only if *X* is in the state that occupies the causal role of pain for the *appropriate* population. (ibid., p. 218)

This is somewhat ambiguous. It is essential to realize that Lewis is relativizing **physical the state that occupies the causal role of pain** (or as we have put it, the realization base) to the population, not the causal role (system of inputs and outputs) itself. The system of inputs and outputs is the same for pain in every possible world, but the state that occupies that causal role is multiply realizable. Once we realize this, then we will see that the mad person would count as being in pain by our formula, as long as we relativize the relationship of being a *realization base* to the appropriate population:

> There exists a set of causal inputs and outputs *C*, such that it is necessarily the case that for all events *x*, *x* instantiates *pain* if and only if there exists a physical property P such that *P* is a *realization base in the appropriate population* for *C* and *x* instantiates *P*.
>
> $\exists C[]\forall x$ (pain(x) $\leftrightarrow \exists P$(realization base in the appropriate population(P,C) $\vee$ P(x)))

So the mad person is undergoing the same realization base as other (appropriate) people do, it is just the case that his wires are hooked up wrong.

As Lewis admits, understanding this requires a coherent notion of the appropriate population.

> But what is the appropriate population? Perhaps (1) it should be *us*; after all, it's our concept and our word, (2) it should be a population that *X* [the object undergoing the event of being in pain in the above formula] himself belongs to, and (3) it should preferably be one in which *X* is not exceptional. Either way, (4) an appropriate population should be a natural kind---a species, perhaps. (ibid., pp. 219-220)

Lewis goes on to argue that, for non-mad people we should use criteria (1), (2), (3), and (4), for mad people we should use criteria (1), (2), and (4), for Martians we should use criteria (2), (3), and (4), and for mad Martians we should use criteria (2) and (4).

Lewis considers the possible counterexample of a creature that is mad, alien, and unique and asserts that such a creature is not possible. On the other hand, he does not consider other fairly obvious counterexamples.

Lewis' position is inconsistent with the logical possibility of a species of mad aliens who felt pain that was both physically realized differently than our pain and which occupied a very different causal role.

His view is also inconsistent with the logical possibility of a species of robots who occupy the same causal role as we do when we are in pain, whose state that realizes that causal role is very different than ours, and who do not feel pain.

One might have a strong intuition that these robots are possible in the weak sense that *for all we know* such robots who are functionally indiscernable from us, yet lacking in mental life, might be created. However, if they are possible in this weak sense that we can't

know that they are impossible, then we can't claim to know that Lewis style functionalism is true, since Lewis style functionalism transparently entails that such robots are not possible. To repeat:

> (1) Lewis style functionalism straightforwardly entails that robots who are functionally indiscernable from us, yet lacking in mental life, cannot exist.
>
> (2) Thus if we know that Lewis style functionalism is true, (given that we know 1), it follows that we can know that robots who are functionally indiscernable from us, yet lacking in mental life, cannot exist.
>
> (3) But for all we know, robots who are functionally indiscernable from us, yet lacking in mental life, can exist.

Therefore, we don't (and possibly can't) know that Lewis style functionalism is true.

Very often in philosophy epistemic humility and semantic humility are in a sort of inverse relationship with one another. For example, suppose the state that the plays the causal role of pain for the vast majority of people is C-Fibers firing; and also suppose that for some people C-Fibers firing plays the causal role of thirst. For Lewis there would be no fact of the matter concerning whether or not these people are in pain. Referring to the above criteria he writes,

> Criterion (1) suggests calling his state pain and regarding him as an exception; criteria (2) and (3) suggest shifting to a subpopulation and calling his state thirst. Criterion (4) could go either way, since mankind and the exceptional subpopulation may both be natural kinds. (Perhaps it is relevant to ask whether membership in the subpopulation is hereditary.) (ibid., p. 220)

There is a kind of semantic humility at work here. Lewis is pointing out that our concepts are not such that they have determinate extensions at every possible world.

Likewise Lewis wants to argue that there may be no fact of the matter about many other philosophical *Gedanken*experiments.

> The interchange of pain and thirst parallels the traditional problem of inverted spectra. I have suggested that there is no determinate fact of the matter about whether the victim of interchange undergoes pain or thirst. I think this conclusion accords well with the fact that there seems to be no persuasive solution one way or the other to the old problem of inverted spectra. I would say that there is a good sense in which the alleged victim of inverted spectra sees red when he looks at grass: he is in a state that occupies the role of seeing red for mankind in general. And there is an equally good sense in which he sees green: he is in a state that occupies the role of seeing green for him, and for a small subpopulation of which he is an unexceptional member and which has some claim to be regarded as a natural kind.

> you are right to say either, though not in the same breath. Need more be said? (ibid., p. 220)

One might be sympathetic to this kind of semantic humility generally, but still have intuitions that go against it in the case of pain. What more to be said might be that necessarily being in pain involves a hurting feeling. If this seems to obviously counterexemplify Lewis' theory perhaps it is just because he tackles such a difficult problem. For it must be noted that functional state identity theory also entails that functionally identical robots necessarily feel pain.

In addition, like Ramsey sentence functionalism, causal role functionalism is not clearly inconsistent with functional state identity theory. We will return to this difficult issue when we discuss the generation problem in section 5a.i.

d. Teleological Functionalism

All three varieties of functionalism discussed above try to accommodate multiple realizability in a way consistent with the hardware/software analogy. Patricia Churchland criticizes the analogy as being potentially very misleading though.

> One powerful objection, repeatedly raised but never answered by those who live by the software analogy, is that the conceptual distinction between hardware and software does not correspond to any real distinction in nervous systems (note: See Churchland and Sejnowski 1992 and Bell 1999). There are many levels of brain organization, ranging from protein channels in membranes, to neurons, microcircuits, macrocircuits, subsystems, and systems. . . At many brain levels there are operations fairly describable as computations, and *none* of these levels can be singled out as *the* hardware level. For example, computations are performed by parts of dendrites, as well as by whole neurons, as well as by networks of neurons. Learning and memory for example, involve computational operations at many levels of structural organization. . . The fact is, in nervous systems there are no levels of brain organization identifiable as *the* software level or *the* hardware level. Consequently, the linchpin analogy (mind/brain = software/hardware) is about as accurate as saying that the mind is like a fire or the mind is like a rich tapestry. (Churchland, (2002, p. 26)).

She is then able to use this to argue against philosophers' attempts to use the computational metaphor to separate cognitive psychology (the study of the software) from neuroscience (the study of the hardware).

> Notwithstanding the strictures of functionalism, the fact is that neuroscience and cognitive science are coevolving, like it or not. This coevolution is motivated not by ideology, but by the scientific and explanatory rewards derived from the interactions. Increasingly, this trend means that data from neuroscience are having an impact on how we frame questions about the mind and how we rethink how best to characterize psychological phenomena themselves. (Churchland, (2002, p. 27))

This is commonsense to anyone who has researched successive versions of psychiatry's *Diagnostic and Statistical Manual of Mental Disorders*. If Fodor was right, then discoveries in neuroscience and pharmacology could never lead to revision in the *Manual*. But Fodor is clearly wrong in that neuroscience and pharmacology have had strong effects on our cognitive understanding of ourselves. Churchland's point should be familiar from our discussion of the way neurophilosophy improves upon traditional type and toke identity theories (section 3c.) and from our discussion of Kim's criticism of Ramsey sentence functionalism (section 4b.iii.).

Though Churchland leads the way in thus arguing against the use of computational metaphors to bifurcate the study of the mind from the study of the brain, she has not attempted to undermine artificial intelligence's role in cognitive science. Nor has she attempted to undermine a suitably weakened functionalism.

Daniel Dennett (1992, 1998), William Lycan (1999), and Elliot Sober (1999) have all defended the idea that cognition can be thought of in terms of nested functions. For example, vision can be thought of as a set of functions, one of which is to recognize lateral symmetry. Then the manner in which we recognize lateral symmetry can be thought of functionally too in terms of the procedures our brain follows to achieve the recognition. Likewise, the brain itself can be thought of in terms of nested functions. For example, a large area might subserve vision, and the ability to recognize lateral symmetry might localize in a part of this region. When generalized appropriately, this teleogical functionalist view is more plausible precisely because it is consistent with the known facts of the brain to which Churchland calls attention, and because it renders cognitive psychology and neuroscience interdependent.

Interestingly, such hybrid views also have the result of making us think of computation differently. In *The Engine of Reason, The Seat of the Soul* Paul Churchland uses the structure of neurons and results about brain function to argue for connectionist, or neural network, models of mental computation, which (he argues) are much closer to how the brain actually works than Von Neuman computers. Then one of the main tasks of cognitive science is that described by Dennett (Dennett, 1992). How does a system that computes in the manner the brain does (one better described with connectionist systems) give rise to rule governed computations that seem to be involved in linear reasoning and language use (abilities better described with languages like LISP running on standard digital computers). Recently this has led to exciting work (Sun, 2001) by computer scientists attempting to develop programs that model this ability.

In this manner, as with our discussion of logical positivism and neuroscience, we see the end result of computation for the philosophy of mind not being an entire philosophy, but rather part of an increasing ecumenicalism. From what has been said in this section and the previous two, it is clear how Dennett and the Churchlands can be thought of as responding to the problems with one-sided philosophies of mind by carefully crafting subtle views that borrow from all three traditions. At the beginning of the twenty-first century, they have argued most vigorously that (and done the most to show how) the

scientific and manifest images can be reconciled. This being said, we are nowhere close to solving all of the problems stemming from Descartes' worry. We now turn to a set of these .

5. Forty-Six Open Problems

I have organized this article in terms of the dialectical progress of three major paradigms, only raising such issues as are sufficient for moving the dialectic forward. Another principle of organization would be in terms of the issues themselves, showing the contributions each paradigm makes to their solution or dissolution. As both approaches have obvious advantages, it behooves us to close with a discussion of the issues themselves.

Since we have thus far chronicled the progress of the three major twentieth century paradigms we do well to conclude with those issues and approaches we can expect and hope to make the most progress on in the current century.

a. Traditional Problems

Remember the mind's tasks: (a) to represent the world through beliefs, desires, perceptions, feelings, and emotions, (b) to reason about which beliefs and desires are correct, and (c) to initiate action. All of our open problems concern one or more of these three tasks, but three seem so insoluble that they might be regarded as perennial philosophical problems in their own right.

i. The Generation Problem

David Chalmers (1996) divides philosophical problems into two types, easy and hard: the easy ones being those that seem treatable by the above paradigms, and the hard ones being those that resist. Easy problems concern the physical bases of mental tasks that can be described computationally, such as our ability to discriminate different pitches along the twelve-tone scale. The main hard problem is consciousness, and William Seager (1999) has stated this problem in what is perhaps canonical form.

(1.) What is it about the brain that allows it to produce consciousness?

We know that brain activity correlates with consciousness but it is not clear we know why. Seager argues that no answer attempted thus far is satisfactory. In particular, he considers representations of the three paradigms discussed above and finds them all wanting.

An adequate answer to problem (1.) will require great understanding of brain function and consciousness itself. It is hoped that our new century's attempted solutions to problem (1.) will thus provide answers to a number of other questions about the nature of consciousness.

(2.) Are robots capable of consciousness? Why or why not?

The thought is that if we truly understand what it is about human brains that make consciousness, then we would have a criterion for consciousness that would be applicable in new and counterfactual situations.

If problem (2.) seems fanciful, consider the following.

(3.) To what extent are animals, birds, fishes, etc. conscious, and how is their consciousness like or different from ours?

An adequate understanding of the neural basis of our own consciousness should shed light on the mental life of other creatures with brains. In this regard, it is interesting to note that Thomas Nagel (1974) wonders whether in principle we cannot "know what it is like" to be a bat, a creature that uses sonar to perceive the world. Serious work on the generation problem, accompanied by progress in animal biology and ethology will help us address this problem.

ii. The Problem of the External World

A significant part of the Cartesian revolution in philosophy was Descartes' characterization of mental acts as representations. That is, since our thoughts are about states of affairs in the world in a manner similar to paintings being about their subject matter, our thoughts represent reality. But then a number of questions immediately arise about the relationship between our perceptions and the external world that they represent, perhaps the main problem being the following.

(4.) How successfully do our perceptions and beliefs mirror the external world?

Skeptics answer "not very," realists answer "very," and idealists answer "not at all," because they deny the existence of the external world. In early modern philosophy this tradition of thought culminated in Kant, who bifurcated reality into the phenomenal (about which one could be an "empirical realist") and the noumenal (about which one could be a "transcendental idealist") realms.

The contemporary philosopher who has done most to extend this tradition is Crispin Wright ((Wright, (1994, 2003)), (Haldane & Wright, (1993))), who provides a typography of possible debates about objectivity in terms of how to characterize truth. Wright realizes that this is a semantic rather than epistemological approach to the worry voiced in (4.), but hopes that such an approach will lead to new insights.

It is one thing to hold that our perceptions and beliefs are largely correct, and quite another to hold that an informative philosophical theory of this correctness is in the offing. Thus:

(5.) Can an informative theory about the relationship between our perceptions, beliefs and the external world be given?

Crispin Wright (1994) calls the combination of a positive answer to (4.) and a negative one to (5.) "quietism," and traces its canonical formulation to the late works of Ludwig Wittgenstein (1972, 2002). More recently, Richard Rorty (1981) and John McDowell (1996) have given influential defenses of quietist positions.

While Rorty might answer a qualified "very" to (4.), he does argue trenchantly that representational metaphors going at least as far back as Plato are mistaken. So there is a sense in which Rorty would like dissolve problem (4.) as well. This brings us to the following.

(6.) If the answer to question (5.) is "no," which of the following is best motivated: skepticism, quietism, or dialetheism?

We will return to skepticism in 5e.ii. Dialetheism is the view that there are true contradictions. While it is anathema to most philosophers, Graham Priest has produced a series of publications arguing powerfully for it. In his recent book, *Beyond the Limits of Thought*, he considers traditional debates about the possibility of a theory of representation to motivate this belief. So the correct answer to both problems (4.) and (5.) might be "yes and no."

iii. Free Will

If free will is an illusion, it is surely a necessary one. Arguably, without the assumption that we could do otherwise there would be no point in rational planning of our actions or in holding ourselves and others morally accountable.

All the more reason why it is an embarrassment that we don't really know what free will is. The ability to have done otherwise seems to presuppose some indeterminism in the world, but indeterminism is surely not sufficient. The mere fact of us doing some things by chance does not render us responsible for those things. For example suppose a cruel dictator rolled dice and forced us to act based on the role of the dice. Such acts are indeterminate, yet not free. Now imagine that the dictator's victim's brain included the equivalent of a dice roller, causing her to do the same acts. Most of us have the intuition that this is still not enough. Therefore, one major conceptual question is the following.

(7.) Given that indeterminism is not sufficient for free will, what else is needed?

An answer to problem (7.) is the major prerequisite to answering the big question.

(8.) Does free will exist?

One should immediately wonder about philosophical and practical results of a possible negative answer to problem (8.). We will take this up again in section 5c.v.

Of course we should ask all of these in an appropriately Cartesian manner, consulting the relevant scientific and philosophical results. Of all contemporary philosophers Robert Kane (1998) has done the most to initiate this dialogue.

b. Issues Overlapping Psychology and Linguistics

Two developments in the last half of the twentieth century have opened up a wide set of new issues for philosophical reflection. The first was the anti-behaviorist revolution, discussed above, championed by Chomsky. License to think about what is going on in Skinner's "black box" has revolutionized psychology. In particular, psychologists now freely theorize about concepts, trying to come up with a model that explains human classificatory behavior. Strikingly, this is a model of things thought to be, in some sense, "in the head;" it is by grasping the concept of a carburetor that I am able to differentiate carburetors from non-carburetors and to have true and false beliefs about carburetors.

The revolution in linguistics is oddly both Chomskyan and anti-Chomskyan. It was Chomsky who first popularized the idea that a grammar should be generative, in that its syntactic component should consist of mechanisms that put words and phrases together to generate all and only the sentences of a language, and that its semantic component should show how facets of the meaning of a sentence are functions of the meanings of the sentence's parts and the way those parts are put together. In early generative linguistics Chomskyans accepted Church's Thesis, equating the generative with the recursive/Turing machine computable. This led to the field of computational linguistics

In linguistics, orthodox Chomskyans now both reject Church's Thesis and the need for a generative semantic theory. The reasons for this are somewhat complicated. However, two facts are relevant for understanding the deeper issue. First, computationally tractable approaches to syntax (e.g. Head Driven Phrase Structure Grammar, Categorial Grammar, Tree Adjoining Grammar) do not utilize transformations. Rather they work with phrase structure rules and add structure to the lexicon. It was by using one of these non-Chomskyan generative syntaxes that Montague was first able to recursively correlate a non-trivial fragment of English with an interpreted formal language, producing the first example of semantics that can truly be claimed to be generative. Second, the attempt to do semantics within a transformational framework (called "Generative Semantics") failed utterly after more than a decade's hard work by a generation of the world's brightest linguists. While Chomskyans viewed this failure as casting aspersions on semantics itself, computational linguists and Montague semanticists saw the failure as resulting from the use of transformations combined with Chomskyans commitment to letting a now controversial theory of language learning (see section 2b.iii. above) affect the structure of a syntax more than the distributional judgments which a syntax is supposed to explain. Recent work (Johnson & Lappin, 1997, 1998) has verified these suspicions to the satisfaction of computational linguists.

While the philosophy of science issues in the "linguistics wars" are fascinating in their own right (Ney, 1957), there is much relevance for the philosophy of mind as well. The

advent of truly generative syntax and semantics has turned much of what used to be solely in the realm of philosophers (philosophical logic) into a natural science. This is so much so that one could argue that philosophers of language ignorant of the role the lambda calculus plays in securing a generative syntax-semantics interface are like philosophers of nature ignorant of the role the calculus played in explaining acceleration in the decades after Newton. This being said, (and as I will argue) there is tremendous opportunity for new philosophy here, done by philosophers immersed both in traditional philosophy and the relevant revolutionary work in psychology and linguistics.

i. Concepts/Word Meanings

Since Frege's groundbreaking work in logic and the philosophy of language, philosophers have come to widely accept three ideals for meanings/concepts.

> (a) All and only linguistic units with the same meaning are intersubstitutable in all contexts (e.g. "the morning star" and "Venus" have different meanings, as "Frank believes the morning star is beautiful" can be true, while "Frank believes Venus is beautiful" simultaneously false, for example, when Frank doesn't know that the morning star and Venus are the same thing).
>
> (b) Knowledge of meaning/possession of concepts explains classificatory behavior (e.g. knowing the meaning of the word "cheese" allows me to distinguish cheese from non-cheese).
>
> (c) The referent of a word depends upon that word's meaning and the way the world is (e.g. the thing I am holding is a pen, but it might not have been a pen if either the word "pen" meant something else, or if the world had turned out differently).

These all seem commonsensical, so it is surprising that the first question one must ask is

> (9.) Does anything exist that satisfies *desiderata* (a), (b), and (c)? If not, must our notion of word meaning bifurcate?

Ever since Hilary Putnam (1975) argued that nothing could satisfy both (b) and (c) a tremendous literature has arisen over problem (9.), and there is currently no widely accepted solution.

Given that concepts are supposed to do so many different things, it should not be surprising that there are so many different models of them. The classical model explains concepts as definitions, or necessary and sufficient conditions. The prototype model explains them in terms of weighted clusters of properties stereotypically associated with the concept. The exemplar model explains them in terms of similarity to paradigm cases of the concept. The theory-theory explains them in terms of roles in theories. Verificationist accounts explain them in terms of practical abilities held by those who possess the concepts. Possible world accounts explain them in terms of the extension of the concept across a set of possible worlds. Again, however, at the beginning of the

twenty-first century there is no unanimity among psychologists, linguists and philosophers about the following.

(10.) What are concepts explained in terms of?

The best current anthology on different models of concepts is Margolis and Laurence's *Concepts: Core Readings*. Given that each theory explains some aspect of human cognitive and classificatory behavior, it is hoped that some day a unifying account will be discovered.

ii. Content/Sentence Meaning

In addition to (a) through (c), concepts/word meanings are almost universally thought to have the following property.

(d) The meaning (or content) of a sentence is the result of combining the meanings (concepts) of the words that compose that sentence.

Currently only the accounts of contents as definitions and as extensions at possible worlds can be fairly argued to uphold (d). In an influential paper, Ernest Lepore and Jerry Fodor (1996) argue that approaches such as the prototype theory are doomed to contradict (d). At this point, one of the most important issues in the theory of concepts is therefore the following.

(11.) How does a theory of concepts interact with a theory of content?

One might conclude a similar thing from Lepore and Fodor's article as many have concluded from Putnam's Twin Earth argument. Perhaps nothing can do the job of (a) through (d). For example, perhaps prototype theory will explain (b) classificatory behavior while a possible worlds account of content will explain (d) the compositionality of concepts. Given how willing many are to bifurcate contents between the classificatory function (b) and the semantic function (c), it is strange that no one has responded in this manner to Fodor and Lepore.

Exciting recent research on (11.) has come out of the work of Ray Jackendoff (1992) and James Pustejovsky (1998). Analogous to non-transformational syntacticians who add more syntactic structure to lexical entrees, Jackendoff and Pustejovsky have both put forward influential proposals about adding more semantic information to the lexicon than in traditional Montague Grammar.

Since, by (d), contents are supposed to be the meaning of sentences, it is easy to get analogs to (a) through (c) above. Since (a) just mentions linguistic units, it can be accepted as is, while the epistemic principle (b) and semantic principle (c) can be recast in the following way.

(b') Knowledge of meaning/possession of contents explains classificatory behavior (e.g. knowing the meaning of the sentence "Cheese is food," plus knowledge of the relevant states of the world, allows me to determine that "Cheese is food" is true).

(c') The truth-value of a sentence depends upon that sentence's meaning and the way the world is (e.g. the sentence "I am holding is a pen," is true but it might not have been if either the word "pen" meant something else, or if the world had turned out differently).

Thus, we can immediately ask questions analogous to those we asked above. In particular:

(12.) What is content explained in terms of?

Given (d), there are clearly as many different models of content as there are concepts. However, given the changes from (b) and (c) to (b') and (c'), a theory of content has additional explanatory requirements. First,

(13.) When are two contents the same, and why? (e.g. "Schnee ist weiss" and "Snow is white" have the same content, while "Frank believes that Schnee ist weiss" and "Frank believes Snow is white," as well as "3 + 5 = 8" and "2 + 7 = 9," do not.)

(14.) When does one content entail the other and why? (e.g. "Frank painted the red boat" entails that Frank painted a boat, while "Frank and Mary painted the red boat together" does not entail that Frank painted a boat).

To date, work within the tradition of Montague semantics is alone in even coming close to answering these questions, albeit it is widely (though not universally, e.g. (Stalnaker, 1981), (Dennett, 1998)) accepted that a possible worlds account of content is fails to explain when sentences with propositional attitude verbs such as "believes that" imply one another. Newer "structured propositions" (Cresswell, 1985) accounts of content are consistent with the general Montagovian paradigm. However, as with the theory of concepts, there is a general sense that each theory is getting something right, and what is needed is a unifying account that explains the successes and failures of each.

iii. Scope and Epistemic Status of Linguaform Explanation

Problems (9.) and (11.) are closely connected to one of the outstanding contemporary issues in artificial intelligence. For assume that there is a unitary model of concepts and that they, of necessity, occur in contents, which are explicated sententially. Then it seems that cognition requires language, if not a spoken one then minimally a language of thought. Then the role of artificial intelligence is to make explicit this implicit language and show the role it plays in our psychical (and physical) economy. More broadly, we can ask:

(15.) Under what conditions are linguaform explanations appropriate in cognitive science?

The hyper-cognitivist insists on linguaform explanations all the way down. So, for example, practical abilities such as riding a bicycle or throwing a baseball are to be understood in terms of the brain issuing specific instructions to the muscles, instructions that can be made explicit via a computer program. The anti-cognitivist will argue that this puts the cart before the horse. Rather, linguaform abilities are in need of explanation in terms of nonlinguistic capacities. This is the Dennettian charge, discussed earlier.

If we assume that linguaform explanations have some role in cognitive science, we must then face the next question.

(16.) What is the epistemic status of linguaform explanations? In what sense are they known by those described by them? In what sense is it correct to say that they are known *a priori*?

Again, there is currently no settled consensus view about this issue. Jon Cogburn (forthcoming) attempts to give a neo-behaviorist theory of when linguaform explanations are applicable and when they are not. The stronger view that linguaform explanations go all the way down, plus a suitably strong version of Chomsky's poverty of stimulus argument, entails that we have (following Chomsky and Fodor) innate knowledge of the concept of a carburetor. The most recent substantive contribution to this debate over innateness is Cowie's *What's Within: Nativism Reconsidered*, a powerful critique of Chomsky and Fodor's nativism.

iv. Animal Cognition

The holy grail of debates about concepts, content, and the scope of linguaform explanation is the following.

(17.) To what extent do animals, birds, fishes, etc. have contentful thoughts?

Progress on problem (17.) will help us answer problems (9.)-(16.). Animals do not have language in the sense of generative procedures to express an unlimited number of new thoughts in terms of a finite number of primitives. If, nonetheless, we can characterize animal thinking, then much progress will have been made on the Dennettian quest to explain linguaform capacities in terms of other mental abilities.

v. Emotions

Recent work by Antonio Demasio (1995) on emotions has stunned and surprised the intellectual world. Demasio has made his life's work the study of people with damage to their left prefrontal cortex who have as a result deadened emotions. Surprisingly, many such people score very well on written tests of cognitive abilities, even those concerning human planning. Yet their ability to plan and act rationally in their own lives is severely

hampered. As a result, it seems that there is a much closer tie between emotion and reason than we might have thought. Demasio's work has reinvigorated attempts to address the following two problems.

(18.) What are emotions explained in terms of?

(19.) What are connections between emotions and other mental states?

While there is consensus that emotions have phenomenal, cognitive, functional, and biological aspects, again we have no unifying theory of how these work together.

c. Issues in Broader Philosophy Relevant to the Philosophy of Mind

At several points in our discussion of the three dominant twentieth century philosophy of mind paradigms we touched upon issues that overlap with other areas of philosophy. The logical positivist philosophy of science motivated behaviorism. Type-type physicalism failed in large part because it entailed that cognitive psychology is completely separate from the brain sciences. We saw that both philosophy of science and metaphysics are relevant in knowing what to make of imaginative *gedanken*experiments (Putnam's *x*-worlders, zombies, angels, etc.). Finally, mental states such as being in pain are of incredible import to ethical issues. Thus, further progress in the philosophy of mind requires more attention to the philosophy of science and metaphysics. Conversely ethicists cannot ignore the philosophy of mind.

i. Reduction

In "Special Sciences" Jerry Fodor famously argued that cognitive psychology was irreducible to the brain sciences. Unfortunately, as philosophers of science have pointed out (Wilson, 1982), the model of reduction Fodor used is one in which nothing has ever been reducible to anything else! Although Mark Wilson is one of the most influential living philosophers of science and language, philosophers of mind have largely not caught up with his work, and still nearly universally talk about reduction in terms of logically deriving one set of fundamental laws from another. But, again and for a host of reasons, such reductions have never been performed in the actual sciences. One such reason is that scientific explanations are not logical axiomatizations. Reducibility in the sciences is almost always thus of necessity much less formal than the positivist model.

To the extent that a formal model of reducibility ever applies in the sciences, it does not involve deriving axiomatizations; rather, it involves showing the equivalence of two equations when the value of some variable in one of the equations is either in, or at, some limit.

Thus a hugely pressing problem for philosophy of mind is the following.

(20.) How do models of reducibility actually used in science apply in the philosophy of mind?

A fascinating recent book by Robert Batterman (2001) investigates formal models of explanation and reducibility from the sciences. Batterman's work demands a substantial rethinking of one of the major strands of 20th century philosophy of mind.

ii. Emergence

Metaphysicians characterize a property as secondary if its being is dependent upon the human mind. For example, one (e.g. (Russell, 1998)) might hold that the color we see is a facet of our phenomenal space, not a real property of the world.

Emergent properties are those supposedly not explicable by fundamental science, but which emerge from complex structures of those things treated by fundamental science. For example, Dennett (1992) argues that it is possible that properties like "red" do not correspond to any natural (or even possibly algorithmically specifiable) collection of states describable in the language of physics. Somehow "red" emerges from a (possibly non-algorithmically specifiable) disjunction of such states.

So what makes all red things red? One might hold that this is because all red things look red to us (in appropriate conditions). But this is to say that red is a secondary property. In this manner, one might start to think that all emergent properties are secondary. However, one might also think that mental properties are paradigmatic cases of emergent properties, e.g. instances of the mental property of pain emerging on brain states. This is not problematic in itself, unless one hoped that the analysis of mental properties as emergent was broadly physicalistic. In that case, the secondary nature of emergent properties renders the analysis circular. To reiterate, mental states are emergent on physical states. But all emergent properties are secondary, and such secondary properties depend for their being on the human mind. Therefore:

(21.) Are there any good models of emergence that are not viciously circular when applied to the philosophy of mind?

In (Cogburn & Silcox) the authors show their own analysis of computational emergence to be so circular, and hence use the theory of emergence to argue against the computational theory of mind. The issues relevant to answering (20.) are more general though, and again, Batterman's work is the starting point for debate on this problem.

iii. *Gedanken*experiments

Our discussion has been littered with genuinely weird thought experiments: Putnam's superspartans, Chalmers' zombies, Cartesian angels. In contemporary philosophy of mind, such *gedanken*experiments are legion, and are used to argue for the most non-trivial conclusions. Putnam (1975) argues against a unitary model of word meanings in terms of a world where water has a radically different micro-structure than on earth.

Jackson (1986) argues against functionalism by positing a person who grows up never having seen red yet knowing the complete scientific theory of redness. John Searle (1980) argues against artificial intelligence in terms of a room that discourses in Chinese. Many (see the discussion in (Churchland, 2002)) have postulated the possibility of somebody's entire color spectrum being inverted. Ned Block (1990) discusses the possibility of an inverted earth, where the colors of things themselves are inverted. Cogburn and Cook (forthcoming) extend this to discuss an inverted space, where points are replaced with lines and lines with points. Many students of philosophy find this proliferation of strange possible worlds dizzying.

Philosophers of mind are lucky that first-rate works in broader philosophy such as Batterman's have recently appeared. Roy Sorenson (1998) has provided a theory of thought experiments that attempts to be general enough to allow us to evaluate thought experiments both within and without philosophy. In addition, Michael Depaul and William Ramsey (1999) have edited a very well received book discussing the status of intuition. The broader discussion sheds much light on how intuitions about these thought experiments may or may not be relevant to the pursuit of truth.

Our earlier discussion of Patricia Churchland raised two very important issues in this regard. Where a muscular account of *gedanken*experiments is one that provides clear reasons for accepting some thought experiments as relevant and dispensing with others, we can ask:

> (22.) What is the appropriate scope of *gedanken*experiments in the philosophy of mind? Is a muscular account possible that does not collapse?

That is, can one craft a theory of *gedanken*experiments that renders some of them (not) O.K. without rendering all(/none) of them O.K. More narrowly, in the case of the follower of Patricia Churchland, this becomes the following.

> (23.) If a neo-empiricist account of *gedanken*experiments (one that renders those used in physics meaningful, but banishes fanciful philosophical ones) is possible, then does this render insignificant putative philosophical distinctions in the same manner as logical positivists attempted?

A plausible, neo-empiricist, non-collapsing account would fundamentally alter the philosophy of mind.

iv. Modality

Churchland's diatribe against *gedanken*experiments traded on the fact that the situations described in them are merely possible (at best), or perhaps merely conceivable yet not possible, and perhaps even (at worst) not even conceivable. In addition to the philosophy of science necessary for adequately addressing (22.) and (23.), we see now the importance of metaphysics.

(24.) What does it mean for something to be conceivable?

(25.) When does conceivability entail genuine possibility?

(26.) What form of possibility is relevant to the philosophy of mind?

Steven Yablo's "Is Conceivability a Guide to Possibility?" is a classic of contemporary metaphysics because it has jumpstarted this important debate. A recent anthology edited by Tamar Gendler and John Hawthorne (2002) contains important articles relevant to discerning solutions to these questions.

One more crucial place the metaphysics of modality might impinge upon the philosophy of mind is that of rigid designation. Following Kripke (1982), a name rigidly designates if it denotes the same object in all possible worlds. This is easy to formally represent as long as one utilizes a fixed domain for all possible worlds.

Unfortunately, the notion of rigid designation as it occurs in the philosophy of mind concerns the designation of predicates, and there is no agreed upon formal way to represent this. In standard modal logics, varying predicate extensions across possible worlds show how things might or might have been in those worlds. So a world where the thing I'm currently holding is not a pen (say it's a cucumber) is not one where the word "pen" means "cucumber" but one where I'm holding something different. Thus, in standard modal logic all predicates both rigidly designate (they only pick out things properly in the extension of the predicate at a given world, given the predicates' meanings) and don't rigidly designate (they pick out different classes of things, but this is interpreted as the worlds being different not the meaning of the words shifting).

This is a fundamental issue in the philosophy of logic (see (Etchemendy, 1999) for how it arises in interpreting non-modal systems), but it adds great confusion to understanding the modal claims of David Lewis (1983), who holds that predicates like "pain" do not rigidly designate, in contrast to say, so-called "natural kind" predicates like (possibly) water. Therefore:

(27.) Can one adequately formulate the claim that predicates do or do not rigidly designate?

One might follow Patricia Churchland and ultimately conclude that these are not differences that make a difference. But it would be nice to be clear about the putative difference at least.

v. Moral Relevance of Mentality

Three facets of the human condition explain the three major moral frameworks of western philosophy. We experience pleasure and pain, and utilitarianism explains our moral duties that arise from this. We are rational/autonomous, and Kantian deontology explains our consequent moral duties. We are social, and social contract theory

thematizes this. Since pain/pleasure and rationality/autonomy are paradigmatic mental states, we thus see a very large intersection between philosophy of mind and ethics.

(28.) What is pain? Which creatures feel pain? What follows?

(29.) What is autonomy? Which creatures are autonomous? What follows?

Work by Valerie Hardcastle (2001) on pain, and by Stephen Wise (2002) on autonomy, have fascinating metaphysical and ethical consequences. Both show how rich an area this intersection is.

Finally, free will raises a host of ethical issues such as the following.

(30.) If free will does not exist, how do we keep from lapsing into despair (giving up rational control of our lives and holding one another morally responsible) or bad faith (acting as if we believe in free will, which we reject in our reflective moments)?

Daniel Dennett (2003) and Patricia Churchland (2002) have recently made a spirited attempt to address this issue. In both cases we again see philosophy of mind making fundamental contributions to ethics.

d. Other Scientific Developments

Of course, logical positivism, neuroscience, and computability theory are not the only scientific paradigms relevant to the philosophy of mind. Recently, notable philosophers have utilized evolutionary theory, quantum physics, and other results from computability theory (not discussed in section 4.) in attempts to forward the dialectic.

i. Computability Reprised

From a logician's perspective, perhaps the most interesting fundamental results in computability theory are negative. From Church, Gödel, Turing and others, we are now able to precisely specify many tasks that computers cannot do. This leads to the following question:

(31.) What are the implications of fundamental limitation results such as Gödel's Incompleteness Theorems, the unsolvability of the halting problem, and the undecidability of (dyadic) first order and stronger logics?

The inability of a computer to algorithmically specify a consistent and complete set of axioms for number theory (proven by Gödel) led J. Lucas (1961) to argue controversially that the computational theory of mind is wrong. Cogburn (2002) produces an independent argument to argue that Gödel's theorem entails that the most plausible non-Platonistic philosophy of math is inconsistent with the computational theory of mind. Cogburn and Silcox (forthcoming) have used the fact that no Turing machine can determine for arbitrary Turing machines and inputs whether those machines will halt for

those inputs (i.e. the unsolvability of the halting problem) to characterize emergent properties.

Given the philosophical fascination with negative fundamental results, it is to be expected that these dialects will continue and grow.

ii. Evolutionary Theory

The claim that the human mind possesses procedures that are not algorithmic (in the sense of Turing machine computable) presupposes Church's Thesis (which equates the algorithmic with the procedural) to be false. The works of Lucas and Cogburn attempt to make this case. In light of this work, Dennett's writings on evolutionary theory (especially his recent *Freeedom Evolves*) raise a fascinating question.

(32.) Can "evolutionary algorithms" violate Church's Thesis?

Remember Dennett's example of color. Perhaps the physical properties underlying all and only red things are so wildly disjunctive that no programmed computer could "measure" them correctly. But natural selection has programmed human brains and sensory systems to be able to do it. Maybe natural selection gives rise to biological machines that violate Church's Thesis. If this were the case, then we would perhaps also have a positive answer to problem (21.)

In broader cognitive science Steven Pinker has appealed to evolutionary theory to explain a host of mental abilities. His popular work (1999) has led many philosophers to ask themselves the following question.

(33.) When are evolutionary accounts of mental abilities merely "just so" stories?

The problem is that it seems that one can take any falsehood about people one likes, and tell a convincing evolutionary story of why that falsehood is true. Example- Why are stepfathers less likely to physically abuse stepchildren? Such behaviors would convince their stepchildren's birth mothers that the stepfather was a good provider, convincing her to have more children with him. Therefore, being nicer to stepchildren is selected for. Of course, you can tell just as convincing an evolutionary story about why stepchildren are more likely to get abused. Any mode of explanation so unconstrained is not science (note: this criticism does *not* hold of mainstream evolutionary biology). When popular writers on evolution use a study of one completely unrepresentative fruit-fly species to claim that Victorian sexual morality is natural (see (Johnson, 2003) for an accessible critique of this ideological abuse of biological science), philosophical scrutiny is warranted.

Finally a very pressing current debate concerns the following.

(34.) Does evolutionary theory support skepticism?

If our beliefs about the world are merely the result of selective pressures in our evolutionary *niche*, then is there any reason to think that these beliefs are true? Alvin Plantiga (1993) has famously argued that the answer to (34.) is "yes," though scientifically minded non-skeptics hope (and argue) the opposite.

iii. Quantum Physics

Common sense counsels the following: (a) reality is never irreducibly statistical, i.e. for any given object *o*, non-vague property *P*, and time *T*, either *o* has *P* at *T* or definitely lacks it; (b) reality is not viewer dependent, i.e. mere viewing of objects does not alter their fundamental properties; (c) there is (by definition?) only one universe; and (d) (at least since relativity theory) nothing travels faster than the speed of light. Unfortunately for common sense, quantum physics almost certainly entails that one or more of the above are false.

The dominant interpretation of quantum physics is Niels Bohr's Copenhagen interpretation, which rejects both (a) and (b). Briefly, on this view fundamental properties of very small objects are irreducibly statistical until viewed, at which point they become concrete. So a particle genuinely is 70% here and 30% there until it is viewed in being in one or another places, at which point it is 100% wherever it was viewed (for details, see (Sklar, 1992)). This raises the following question.

> (35.) Does quantum physics support a form of idealism or pan-psychism that will help dissolve the generation problem and the problem of the external world?

As Chalmers (1996) argues brilliantly, many interpretations of quantum physics treat information as an irreducible, fundamental property of the universe. If this is right, then we should not expect to reduce to the mental to the non-mental, as the universe itself is fundamentally mental.

If the universe is fundamentally statistical, then it contains genuine indeterminacy. There is just no fact of the matter uniquely determining where I will find our elusive particle, or (for example, in the case of nuclear decay) when I will find it. So again, we might ask.

> (36.) Does quantum physics help explain non-computable aspects of mental life?

Roger Penrose (1989, 1994) has controversially argued that nerve cells are sensitive to quantum indeterminacy and that this might provide a better model of mentality than the computational theory of mind.

e. Alternative Philosophical Traditions

In our discussion of the three main paradigms of twentieth century philosophy of mind, most of the dialectic was moved forward in the shadow of logical positivism. This worked at a substantive level with the legacy of behaviorism and at a methodological

level with the analytic philosophers' relentless chasing of dialectic- thesis, antithesis, and synthesis all driven by fiendishly clever arguments and counterarguments.

Just as vistas are opened by bringing in new results from different sciences than those focused on in previous philosophy of mind, new insight arises from decoupling philosophy of mind from some of its traditional twentieth century philosophical presuppositions.

i. Anti-Realism

Paul Churchland (1981), Richard Rorty (1965), and Peter Unger (1980) were the first major philosophers in the analytic tradition to raise the possibility that the reason the manifest and scientific images are hard to reconcile is that the manifest image is largely mistaken. This has had the salutary effect noted earlier (section 3c.) while discussing Patricia Churchland, who argued that scientific advances always lead us to see that the previous manifest image was mistaken in some respects.

Clearly one could maintain the more radical visions of the early (Paul) Churchland, Rorty, and Unger, holding the manifest image to be extremely flawed. Making sense of this more extreme view requires discussing similar extreme views from other areas of philosophy, such as ethics and the philosophy of language. Above (section 5a.) we noted how Crispin Wright and John McDowell's work can be cast as a fundamental rethinking of the problem of the external world. Wright and McDowell's contributions come out of Michael Dummett's philosophy of language.

For Dummett, one can be a realist about some discourses (say physics) and an anti-realist about others (say ethics and math). Moreover, there are a variety of ways to be an anti-realist. Minimally, realists about a discourse *D* hold that:

(a) Propositions of *D* are truth apt,

(b) Some propositions of *D* are true,

(c) The truth or falsity of sentences of *D* are independent of us.

The error theorist in ethics (most analogous to early eliminativism) holds (a) and (c) but rejects (b). The classical non-cognitivist (most analogous to Rorty's (1981) claims about mental states and philosophy itself) rejects (a) and (b) and accepts (c). The cultural relativist accepts (a) and (b) and rejects (c), because human cultures determine which propositions of *D* are true (see (Garner, 1994) for a discussion of the ethical views). The Dummettian anti-realist agrees with the relativist about (a) through (c), but rejects (c) for different reasons, like Kant, holding it to be inconsistent with a suitably strong verificationism (for details, see (Cogburn, 1999)). Thus we must ask:

(37.) Is anti-realism motivated in relation to any discourse related to the manifest image?

Given the radical nature of some forms of anti-realism, we must also ask:

(38.) Which, if any, forms of anti-realism are self-refuting?

Suppose that I claim that belief talk deserves an error-theoretic account. Can I claim to truly believe this? This would be a true belief that there are no true beliefs, a paradoxical view at best.

In closing, it needs to be stressed that the dialectics growing from Wright's (1994) characterization of the commitments of realism renders (a) through (c) slightly cartoonish, albeit (36.) and (37.) remain open questions for contemporary, more sophisticated, accounts of realism/anti-realism debates.

ii. Skepticism

If anti-realism seems to impugn the objectivity of a discourse, skepticism presupposes such objectivity. For the skeptic, there is a fact of the matter rendering our claims true or false, but we can't know what that is. The most influential contemporary skeptic in the philosophy of mind is Colin McGinn (2000). Like Plantinga, McGinn uses evolution (and, in his case, other considerations) to support skepticism. However, Plantinga's *modus tollens* is McGinn's *modus ponens*. McGinn raises the possibility that our brains have evolved to be able to ask reasonable questions about the relation of mind and body, but that we are not evolved to be able to answer such questions. When we think of other animals in their ecological niches, our epistemic pretenses do seem vain. So, analogous to (37.) and (38.), we have:

(39.) In what areas is skepticism motivated?

(40.) Is skepticism about the mind self-refuting?

As with anti-realism, only certain forms of skepticism are clearly self-refuting. For example, we can't, in the manner of Sergeant Schultz, consistently claim to know that we know nothing.

iii. Pragmatism

Sidney Morganbessser famously quipped that pragmatism is great in theory, but not in practice. Strictly speaking, pragmatism is a variety of anti-realism, as *per* the definitions in 5e.i. However, since it is a rich philosophical tradition in its own right, it deserves its own section. Like relativists and Dummettian verificationists, pragmatists reject the realist contention that the truth of propositions is completely independent of human capacities and practices (hence the joke). Classical pragmatists in the Jamesian tradition identify truth with usefulness, while Peircean pragmatists identify truth with what we would believe in the ideal limit of virtuous discourse (for details, see (Misak, 1995)).

This tradition has a resonance for philosophy of mind that is perhaps underutilized. For example, in "The Will to Believe," William James argues that when a choice to believe in a proposition is momentous, genuine, forced, and such that reason alone can't decide, it would be irrational not to consult our "passional nature," which includes value judgments. James' example is of someone compelled to be a theist, though a better religious example might be someone who is already a theist picking which religious community to join. The would-be believer could become a Southern Baptist, with its concommitment articles of faith involving the literal truth of the entire Southern Baptist version of the *Bible*, women's subservience, eternal torture of non-Christians, and attacks on homosexuality. On the other hand, the would-be believer could join the Episcopal Church, which lacks all of these commitments and indeed has ordained a gay bishop. How to decide? Both religions have talented theologians who argue that their doctrine is more in line with the core ideas of Christianity. But what are those? James' genius was to argue that in such cases we must first consult our "passional nature." At the very best then (argues the pragmatist), one's decision reflects much more one's ethical rather than ontological beliefs.

Richard Rorty's masterful *Philosophy and the Mirror of Nature*, can be read as an extended meditation on this theme, the primacy of value. Rorty argues that much of traditional philosophy is a misguided attempt to "ground" essentially normative issues with a metaphysical description of reality- but (for Rorty) this is to confuse prescription and description. An example of this (not discussed by Rorty) is metaphysical debate about "autonomy." Possession of autonomy is supposed to explain a variety of intuitions we have concerning when to hold people responsible and when to legally permit them to do things like participate in medical research. The Rortyan would argue that the notion of "autonomy" here is a useless epicycle, what we should be discussing is the normative issues themselves, attempting to come to a rational consensus about shared values.

While Rorty himself takes his Jamesian insight to obviate the need for much traditional philosophy, this is largely due to his ethical relativism. If one accepted the pragmatist primacy of value, but was a realist about normative matters, then a large part of traditional philosophy would be retained, only under a radically new guise. It is in this vein that one might ask:

> (41.) In what areas are pragmatist solutions to philosophical issues about the mind plausible?

Consider how the neo-Jamesian pragmatist might respond to some of our fanciful thought experiments. Is a robot like Star Trek's Data genuinely conscious? If the choice to believe this is momentous, genuine, and forced, and such that metaphysics can't decide- then we must consult our values. How should we treat Data? If our ethical intuitions (as they almost surely would be) were to treat him *as if* he were conscious, then (for the pragmatist) there is nothing left to debate. The substantive normative issues decide the metaphysical issues. Data is conscious.

iv. Materialist Hermeneutics of Suspicion

Friedrich Nietzsche's *The Genealogy of Morals* inaugurated a new method of philosophy. Following suggestions in his earlier *Beyond Good and Evil*, Nietzsche describes a history of the concepts of "good" and "evil." That is, rather than explain why our judgments about good and evil are true, Nietzsche attempts to explain why we hold our beliefs about good and evil.

Given his penchant for the extreme, it is no surprise that Nietzsche's genealogy of morals undermines the moral system it explains. In addition to Marx's writings of the same era, this kind of philosophizing can be considered the "hermeneutics of suspicion," with the suspicion being that if we really understood why we believe what we do, then our beliefs would be quite different.

In the twentieth century Michel Foucault did the most to continue Nietzsche's heritage, attempting a genealogy of the modern notion of the self. Foucault's fascinating *corpus* raises the following questions.

> (42.) To what extent do the writings of Marx, Nietzsche, Foucault, etc. add to the evolutionary understanding of widespread beliefs (in particular belief in central tenets of the manifest image)?
>
> (43.) To what extent does the correct explanation of why people believe the manifest image support some form of anti-realism about the manifest image?

In a more general sense, a genealogical approach can be applied to the dominant beliefs of philosophers. Why did the computational theory of mind seem so attractive to philosophers of mind in the post World War II era? Is it any accident that many of the most influential thinkers in that era had been involved in logistical problems arising from total war? Does the post World War II rise of the "general manager," trained in business schools in "management," partially explain the penchant for domain general planning and learning algorithms among early artificial intelligence practitioners? Does it explain the ideology of the brain as a central planner, as if the brain were a Harvard trained chief executive officer?

Philosophers tend to eschew such issues, either viewing them as non-philosophical or as tacitly committing the genetic fallacy, though there have been recent sustained attacks on the genetic fallacy fallacy (e.g. (Garner, 1994)), which is the mistake of thinking that the sources of ideas are *never* relevant to assessing their truth. Such dialectics open the door for Foucault inspired philosophy of mind.

v. Phenomenology

If there is any substantive difference between "analytic" and "continental" philosophy, it concerns the reception of Heidegger by French philosophers in the post World War II period. As Luc Ferry and Alain Renault (1990) argue persuasively, the "French Philosophers of the 60s" all had Heidegger in common as philosophical father.

When pressed, "analytic" philosophers will admit that Heidegger's membership in the Nazi party and later (at the very least) bizarre refusal to say the Holocaust was wrong are all irrelevant to questions of the value of his philosophical work. Yet these allegations, when combined with the purported unclarity of some of his writing, worked to hinder Heidegger's reception in the United States. The thought seems to be that purposeful obscurity fits well the personal vices that led him to be implicated in Naziism. Thus some (for a recent example, see Glover (2001)) openly suspect that his *oeuvre* contains much mud and little real depth. That is, perhaps Heidegger's defenders commit the genetic fallacy fallacy.

Two signature events have radically altered this landscape in analytic philosophy. First was the publication of Hubert Dreyfus' influential *What Computers Still Can't Do*. Though Dreyfus had been arguing for years that a proper understanding of Heidegger showed there to be serious problems with then dominant approaches to artificial intelligence as well as the computational theory of mind, with the publication of this book it became clear to the vast majority of analytical philosophers that Dreyfus had indeed (using Heidegger) predicted the main problems that had since clearly beset rule-based approaches to artificial intelligence.

For Dreyfus, standard artificial intelligence's relatively linguaform explanation of mentality confuses *explanans* (that which is used to explain) and *explanandum* (that which needs to be explained). Instead of explaining one's ability to pick up a glass in terms of an explicit language (either computer program or language of thought) as standard artificial intelligence would have it, language and linguaform abilities (such as planning and problem solving) need to be explained in terms of whatever allows one to pick up the glass. For the Heideggerian, as with the pragmatist, first comes practical interaction with the world. Abstract thinking must be explained in terms of this. As noted earlier, to some extent Daniel Dennett's theory of content can be seen as an instance of this framework (albeit perhaps not his explanation of consciousness, see (Seager, 1999)). Dreyfus' fascinating work convinced many analytic philosophers that Heidegger deserves an important place in the canon.

The second main event for the changing reception of phenomenology has been the English language dissemination of the work of members of Fransesco Varela's research group. To Dreyfus' phenomelogical critique of standard cognitive science, Varela added a positive research program. While most analytic philosophers had seen Heidegger's thought as leading up to the kind of paradoxical anti-metaphysics of Jacques Derrida, Varela constructed an alternative history of phenomenology that championed the contributions of Merleau-Ponty.

One critique of early Husserlian phenomenology is that the "phenomenological method" of bracketing all presuppositions and attending to the phenomena itself is impossible: (1) our presuppositions are always there, and (2) psychological testing has shown that people misreport their own mental life. Heidegger himself made the first criticism against

Husserl (see (Moran, 2000)); Paul Churchland (1988) presents the clearest statement of the second.

Rather than retreating to Derridean hermeneutic phenomenology, Varela argues that: (1) an appropriate form of naturalism can correct for the problem of presuppositions, and (2) test subjects can, in a non-question begging way, be suitably trained.

It is impossible to convey the richness and excitement of the work in the naturalized phenomenology tradition. One excellent recent anthology is (Petit, *et. al*., 2000). At this point however, enough is on the table to allow us to pose the following question.

> (44.) Can a "naturalized" phenomenology reinstate introspectionist methods in psychology?
>
> (45.) Can a phenomenological approach to cognitive science avoid traditional pitfalls: (a) badly working A.I, and (b) over-reliance on innateness hypotheses.

In addition to the above prospects, part of what secures the interest in phenomenology from Hegel to Varela is the possibility of dissolving the problem of the external world and the generation problem. Thus:

> (46.) Does the attempted dissolution of the generation problem and the problem of the external world characteristic of phenomenology (of either the neo-Hegelian or neo-Pontyan) work, or is it simply question-begging?

If the world as experienced is fundamental, then the hope is that the disconnect between the world (both outside the body and inside the head) and our experience cannot arise. In this manner, perhaps the deepest of our Cartesian worries will not be solved, but rather dissolved.

References:

American Psychiatric Association, 2000, *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR (Text Revision)*.

Batterman, R., 2001*, The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence,* Oxford University Press.

Bell, A., 1999, "Levels and Loops: The Future of Artificial Intelligence and Neuroscience," *Philosophical Transactions of the Royal Society of London*, B 354, 2013-2020.

Block, N., and Fodor, J., 1972, "What Psychological States Are Not," *Philosophical Review*, 81, 159-181, reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980).

Block, N., 1990, "Inverted Earth," in *Philosophical Perspectives 4, Action Theory and Philosophy of Mind,* 53-79.

Boolos, G., Burgess, J., and Jeffrey, R., 2002, *Computability and Logic*, Cambridge University Press.

Brandom, R., 2001, *Articulating Reasons: An Introduction to Inferentialism*, Harvard University Press.

Brooks, D., 1994, "How to Perform a Reduction," *Philosophy and Phenomenological Research* 54: 803-14.

Calvin, W., 1990, *The Cerebral Symphony, Seashore Reflections on the Structure of Consciousness*, Bantam Books.

Carpenter, B., 1998, *Type Logical Semantics*, MIT Press.

Chalmers, D., 1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press.

Chisholm, R., 1957, *Perceiving*, Cornell University Press.

Chomsky, N., 1966, *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*, Harper and Row.

Chomsky, N., 1959, "A Review of B.F. Skinner's Verbal Behavior," *Language*, 35, reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980).

Chomsky, N., 1995, *The Minimalist Program*, MIT Press.

Chomsky, N., 1996, "Language and the Problems of Knowledge," in ed. A. Martinich, "The Philosophy of Language," (Oxford University Press).

Churchland, P., 1981, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy*, 78.

Churchland, P., 1988, *Matter and Consciousness - Revised Edition: A Contemporary Introduction to the Philosophy of Mind*, Dimensions.

Churchland, P.S., 1986, *Neurophilosophy: Toward a Unified Science of the Mind-Brain*, MIT Press.

Churchland, P.S., 2002, *Brain-Wise: Studies in Neurophilosophy*, MIT Press.

Churchland, P.S., and Sejnowski, T., 1988, "Perspectives in Cognitive Neuroscience," *Science*, 242, 741-745.

Clark, A., 1998, *Being There*, MIT Press.

Cogburn, J., 1999, *Slouching Towards Vienna: Michael Dummett and the Epistemology of Language*, dissertation, Ohio State University.

Cogburn, J., 2002, "Deconstructing Dummett's Anti-Realism: A New Argument Against Church's Thesis," *The Logica Yearbook*.

Cogburn, J., forthcoming, "Inferentialism and Tacit Knowledge," *Behavior and Philosophy*.

Cogburn, J. and Cook, R., forthcoming, "Inverted Space: Minimal Verificationism, Propositional Attitudes, and Compositionality," *Philosophia*.

Cogburn, J. and Silcox, M., forthcoming, "Computing Machinery and Emergence: The Metaphysics and Aesthetics of Video Games," *Minds and Machines*.

Coffa, J. 1991, *The Semantic Tradition from Kant to Carnap: To the Vienna Station*, Cambridge University Press.

Cowie, F., 1999, *What's Within? Nativism Reconsidered*, Oxford University Press.

Cresswell, M., 1985, *Structured Meanings*, MIT Press.

Davis, M., 2000, *The Universal Computer: The Road from Leibniz to Turing*, W.W. Norton & Company.

Demasio, A., 1995, *Descartes' Error: Emotion, Reason, and the Human Brain*, Quill.

Dennett, D., 1989, *The Intentional Stance*, MIT Press.

Dennett, D., 1992, *Consciousness Explained*, Little, Brown and Co.

Dennett, D., 1998 *Brainchildren*, (MIT, 1998).

Dennett, D., "Real Patterns," in *Brainchildren* (MIT 1998).

Dennett, D., 2003, *Freedom Evolves*, Viking Books.

Depaul, M., and Ramsey, W., 1999, *Rethinking Intuition*, Rowman and Littlefield.

Descartes, R., in *The Philosophical Writings of Descartes, Volume I*, tr. J. Cottingham, R. Stoothoff, D. Murdoch, (Cambridge University Press, 1980).

Dreyfus, H., 1992, *What Computers Still Can't Do*, The MIT Press.

Eliot, T.S., 1950, *The Complete Poems and Plays*, Harcourt Brace Jovanovich.

Etchemendy, J., 1999, *The Concept of Logical Consequence*, CSLI Publications.

Ferry, L., and Renault, A., 1990, *French Philosophy of the 60's: An Essay on Anti-Humanism*, University of Massachusetts Press.

Finger, S., 1994, *Origins of Neouroscience: A History of Explorations into Brain Function*, Oxford University Press.

Fodor, J., 1980, *The Language of Thought*, Harvard University Press.

Fodor, J., 1974, "Special Sciences," *Synthese* 2, 97-115.

Foucault, M., 1984. *The Foucault Reader*, ed. P. Rabinow, Pantheon Books.

Friedman, M., 1999, *Reconsidering Logical Positivism*, Cambridge University Press.

Garner, R., 1994, *Beyond Morality*, Temple University Press.

Glover, J., 2001, *Humanity: A Moral History of the Twentieth Century*, Yale University Press.

Haldane, J., and Wright, C., 1993, *Reality, Representation and Projection*, Oxford University Press.

Hardcastle, V., 2001, *The Myth of Pain*, Bradford Books.

Hempel, C., "The Logical Analysis of Psychology." In *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980)

Jackson, F., 1986, "What Mary Didn't Know," *Journal of Philosophy* 83, 291-295.

James, W., 1896, "The Will to Believe, *New World*.

Johnson, D. and Lappin, S., 1997, "A Critique of the Minimalist Program," *Linguistics and Philosophy* 20, 229-271.

Johnson, D. and Lappin, S., 1998, *Local Constraints vs. Economy*, Cambridge University Press.

Johnson, O., 2003, *Dr. Tatiana's Sex Advice to All Creation*, Owl Books.

Kane, R., 1998, *The Significance of Free Will*, Oxford University Press.

Kim, J., 1998, *Philosophy of Mind*, Westview Press.

Kripke, S., 1982, *Naming and Necessity*, Harvard University Press.

Lepore, E., and Fodor., J., 1996, "The Pet Fish and The Red Herring: Why Concepts Aren't Prototypes," *Cognition* 58, 243-276.

Lewis, D., 1980, "Mad Pain and Martian Pain," in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980a).

Lucas, J., 1961. "Mind, Machines, and Gödel," *Philosophy* 36, 112-137.

Lycan, W., "The Continuity of Levels of Nature," in *Mind and Cognition: An Anthology*, ed. William Lycan (Blackwell, 1999).

Margolis, E. and Laurence, S.,1999, *Concepts: Core Readings*, Bradford Books.

McDowell, J., 1996, *Mind and World*, Harvard University Press.

McGinn, C., 2000, *The Mysterious Flame: Conscious Minds in a Material World*, Basic Books.

Merleau-Ponty, M., 2002, *Phenomenology of Perception*, tr. C. Smith, Routledge.

Misak, C., 1995, *Verificationism*, Routledge.

Moran, D., 2000, *Introduction to Phenomenology*, Routledge.

Nagel, T., 1974, "What Is It Like To Be a Bat?" *The Philosophical Review*, LXXXIII, 435-50.

Ney, J., 1993, "On Generativity: The History of a Notion that Never Was," *Historiographia Linguistica* 20, 441-454.

Nietzsche, F., 1989, *On The Genealogy of Morals*, tr. Walter Kaufmann, Vintage Books U.S.A.

Nietzsche, F., 1989, *Beyond Good and Evil*, tr. Walter Kaufmann, Vintage Books U.S.A.

Penrose, R., 1989, *The Emperor's New Mind*, Oxford University Press.

Penrose, R., 1994, *Shadows of the Mind*, Oxford University Press.

Petitot, J., Varela, F., Pachoud, B., and Roy J., 2000, *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, Stanford University Press.

Pinker, S., 1999, *How the Mind Works*, W.W. Norton and Company.

Plantinga, A., 1993, *Warrant and Proper Function*, Oxford University Press.

Pollard, C. and Sag, I., 1994, *Head Driven Phrase Structure Grammar*, University of Chicago Press.

Priest, G., 2003, *Beyond the Limits of Thought*, Oxford University Press.

Putnam, H., "Psychological Predicates," in *Art, Mind, and Religion* , ed. W.H. Capitan and D.D. Merrill (University of Pittsburgh, 1967).

Putnam, H., "The Meaning of 'Meaning'," in *Language, Mind and Knowledge*, ed. K. Gunderson, (University of Minnesota, 1975).

Putnam, H., 1965, "Brains and Behavior," in *Analytical Philosophy*, vol. 2 (Blackwell), reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980a).

Putnam, H., 1967, "The Nature of Mental States," in *Art, Mind, and Religion* (University of Pittsburgh Press), pp. 37-48, reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980b).

Quine, W., 1980, "Two Dogmas of Empiricism," in *From a Logical Point of View*, (Harvard University Press).

Rorty, R., 1965, "Mind-body Identity, Privacy and Categories," *The Review of Metaphysics*, XIX.

Rorty, R., 1981, *Philosophy and the Mirror of Nature*, Princeton University Press.

Russell, B., 1998, *The Problems of Philosophy*, Oxford University Press.

Seager, W., 1999, *Theories of Consciousness: An Introduction and Assessment*, Routledge.

Searle, J., 1980, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, 417-424.

Shapiro, S., 2000, *Foundations Without Foundationalism: The Case for Second Order Logic*, Oxford University Press.

Skinner, B., 1957, *Verbal Behavior*, Appleton-Century-Crofts.

Skinner, B., 1953, *Science and Human Behavior*, Macmillan, reprinted in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Harvard University Press, 1980).

Sklar, L., 1992, *Philosophy of Physics*, Perseus Books.

Smart, J., 1959, "Sensations and Brain Processes," *Philosophical Review* 68, 141-156.

Sober, E., 1999, "Putting the Function Back Into Functionalism," in *Mind and Cognition: An Anthology*, ed. William Lycan (Blackwell).

Sorenson, R., 1998, *Thought Experiments*, Oxford University Press.

Staddon, J., 2001, *The New Behaviorism: Mind, Mechanism, and Society*, Psychology Press, Taylor and Francis Group,

Stalnaker, R., 1981, "Assertion," in *Radical Pragmatics*, ed Peter Cole, Academic Press.

Strawson, P., 1966, *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*, Routledge.

Sun, R., 2001, *Duality of the Mind: A Bottom Up Approach Toward Cognition*, Lawrence Erlbaum Assoc.

Thagard, P., 1996, *Mind: Introduction to Cognitive Science*, Bradford Books.

Turing, A., 1950, "Computing Machinery and Intelligence," *Mind* 59, 433-460.

Unger, P., 1980, "Skepticism and Nihilism," *Nous*, 517-545.

Wilson, M., 1982, "Predicate Meets Property," *Philosophical Review* 91, 549-589.

Wilson, M., 1982, "What is this thing called "pain"? The Philosophy of Science Behind the Contemporary Debate," *Pacific Philosophical Quarterly* 66, 27-67.

Wise, S., 2002, *Drawing the Line, Science and the Case for Animal Rights*, Perseus Publishing.

Wittgenstein, L., 1972, *On Certainty*, Dimensions.

Wittgenstein, L., 2002, *Philosophical Investigations*, Blackwell.

Wright, C., 1994, *Truth and Objectivity*, Harvard University Press.

Wright, C., 2003, *Saving the Differences: Essays on Themes from Truth and Objectivity*, Harvard University Press.

Yablo, S., 1993, "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53, 1-42.