Michael Huemer
Phil 132

ANSWER TO SEARLE ON THE MIND/BODY PROBLEM

The aim of this paper is to refute John Searle's solution (attempted solution?) to the mind/body problem. To that end, I will first briefly identify the problem; second, explain Searle's solution; and third, show what is wrong with it. The theory I'm referring to is given in chapter 10 of *Intentionality* and chapter 5 of *The Rediscovery of the Mind* (perhaps among other places). Therefore, I assume the reader has read these.

1 The problem stated

Broadly, the mind/body problem is the problem of saying what is the relationship between the mental and the physical or, more specifically, between mental phenomena and brain processes.

For example, some people have said that there is basically no relation (parallelism). Some people have said brain processes cause mental phenomena and that's it (epiphenomenalism). Some have said that mental phenomena and brain processes are events going on in two different substances that interact with each other (Cartesian dualism). Et cetera. Notice that these theories all take the form of statements of how mind and brain are or are not related. Some theories may claim a conceptual connection between mental and physical descriptions (as in logical behaviorism); others may claim an identity; others, a causal connection without a conceptual connection; etc. The philosophical debate becomes a debate among competing theories of said relation. Now, in this paper, I don't care about any of the theories except one, namely, the one Searle adheres to. Nor will I attempt a positive statement of what the relation between mind and brain is; I will only prove that Searle's description of the relation is not correct.

2 Searle's solution

There are at least two different ways of describing a relationship. One would be to describe it directly, explicitly. That is, if there exist adequate words referring to the relationship, then one may simply name it. One may say, "xRy." A second way would be to give an analogy. One may say, "A is to B as C is to D." This means, "The relationship between A and B is the same as the relationship between C and D." If the hearer knows what the latter relationship is, then he can use your statement to see what the former relationship is - namely, the very same one. Notice that there is nothing metaphorical about this. An analogy is a literal statement. If I say, "Man is to boy as woman is to girl," then I am saying that the relationship between a man and a boy is the identical relation as that between a woman and a girl. And this may be either true or false. Also notice that in an analogy I do not have to actually specify the relationship between A and B; I only say that whatever it is, it is the same as the one between C and D.

Searle's solution to the mind/body problem describes the relation between the mind and the brain in both ways. On the one hand, he gives a few analogies to explain the relation; and on the other hand, he explicitly describes what that relationship is. Therefore, I will attack both his analogies and his explicit specifications. It is appropriate to attack analogies in this case because claiming an analogy is claiming something about relationships, and the mind/body problem focuses specifically on a relationship. By showing that the analogies Searle offers fail, I thereby show that his view of the relationship between mind and brain is false, since the analogies are indirect descriptions of the mind/brain relationship, as I explained above. Additionally, I will attack his explicit specifications of the relationship as faulty.

For the sake of brevity, I assume that the reader is familiar with Searle's account; this statement is for purposes of reminding him and also of getting the account more precise and explicit than it already is in Searle's writings.

2.1 The analogies

I'm only going to consider two of Searle's analogies. The application of my arguments to the rest of his analogies as well as to any other ones that might be thought of in the same general vein will be obvious.

The first analogy - call it "the liquid analogy" - says

$$\frac{\text{mental states}}{\text{brain processes}} \simeq \frac{\text{liquidity}}{\text{behavior of } H_2O \text{ molecules}}$$

To be careful, we should say that the behavior of neurons in the brain is to the mental states experienced by the person whose brain it is as the behavior of H_2O molecules in a quantity of water is to liquidity. (This analogy is advanced on p15 of the reader and Chapter 10 of Intentionality.) For brevity, hereafter I shall refer to this case as "the liquid case" and the relation between the liquidity of water and the behavior of H_2O molecules as the "liquid/H_2O relation".

The second analogy - call it "the heat analogy" - says

$$\frac{\text{mental states}}{\text{neurological processes}} = \frac{\text{heat}}{\text{mean molecular kinetic energy}}$$

The relationship between any given mental state and a certain neurological process is the same as the relationship between heat and mean kinetic energy of molecules. (This analogy is suggested throughout section IV, chapter 5 of the reader and in "Minds and Brains without Programs".) Hereafter, I shall use the terms "the heat case" and "the heat/molecular motion relation"; I trust there will be no confusion.

Searle must be presupposing that the heat/molecular motion relation is the same as the liquid/H_2O relation, or else it would be self-contradictory of him to offer both analogies. If the relationships were different, then the mind/brain relationship could be identical with one or the other but not both. Below I will draw a distinction showing an important difference between the two cases and then show why the mind/brain relationship cannot match either case.*

2.2 The explicit description

The explicit description of the mind/brain relation that Searle provides is, given the analogies he makes, also a description (mutatis mutandis) of the heat case and the liquid case.

The account states the following:

(1) Mental phenomena are caused by brain processes.

(2) Mental phenomena are also higher-, macro-level features of the brain.

(3) They are not properties of neurons (no neuron is conscious).

(4) They are not compositional properties: you cannot deduce them just from the properties of neurons and their arrangement.

(5) But you can explain them by the causal interaction of the elements (the elements are the neurons).

(6) They do not have any causal powers that are not explained by the causal interaction of the elements.

(7) They cannot be reduced to physical phenomena. (Note that this point refers to 'ontological reduction': "The antireductionist point of the argument is ontological ..." (p99 of the reader).)

It is fairly clear how #1-6 fit with the liquid analogy. Liquidity is a macro-level feature of a system of H_2O molecules that is caused by the behavior of the molecules. No individual molecule is wet. And liquidity and all its causal powers are explained by the behavior of H_2O molecules. However, it is obscure how #1-6 comport with the heat analogy (is heat caused by motion of molecules? Is it explained by causal interactions?), and I am not going to attempt to make them fit.

---

* Of course, the two relationships could be the same in some respects and different in others. Thus, a general description could be given of the nature of both these relationships that specified what they had in common, while at the same time, a more discriminating analysis shows where they differ. This is what I think in fact is the case: Searle correctly identifies (section 2.2) something that they both have in common, while I identify where they differ (section 3).

Further, #7 does not seem to fit with either analogy. Liquidity and heat are reducible. But Searle has an answer to this (see chapter 5, section IV of the reader). The answer basically says, yes there is this difference, but it doesn't have anything to do with a disanalogy in the ontology of the mind/brain case; it just has to do with a difference in our definitional practices with respect to mental terminology. That is, he says that although heat is reducible and consciousness is not, that isn't because heat bears a different kind of relation to molecular motion than consciousness does to brain processes. It is just because we happen to consider a certain aspect of consciousness - its subjectivity - interesting, whereas we don't find the subjectivity of heat interesting. By this argument, Searle hopes to save the analogy between the heat case and the mind case.

3 Objection

Not surprisingly, I don't find Searle's analogies illuminating or appropriate, and I don't find his attempt to explain away the disanalogy vis-a-vis the irreducibility of consciousness at all convincing.

Let's remind ourselves where we are. I said that the problem is to specify the relationship between the mind and the brain. We have just seen Searle's attempt to explain the nature of said relation both by analogies and by explicit descriptions of the relation. He says that mental states are both caused by and realized in the physiology of the brain in the same way that liquidity is both caused by the behavior of H_2O molecules and realized in the collection of molecules. He frequently asserts that there is nothing mysterious about the mind and that it can be explained in a way familiar to us from the rest of science.

I want to show what is wrong with his view by considering two separate cases. If we have a macro-level feature that is realized in a system of micro-elements, then either it is necessary that a system composed of such elements as those, arranged in such a way as that, has this property; or it is contingent. That is to say, either the description of the micro-level phenomena necessarily implies the presence of the macro-level phenomenon or it does not necessarily imply that. Note that herein, by "necessity" I do not just mean physical or nomological necessity but "necessity in the highest degree," as Kripke puts it - that is, truth in all possible worlds. And "possible", of course, is to be understood in the broad sense of conceptual possibility. What I will show is that either way - whether the connection is necessary or contingent - the case will not be analagous to the mind/brain case, and on either assumption, Searle's seven claims listed above are not all true. The two possibilities - that the connection is necessary and that it is contingent - correspond nicely, respectively, to the two analogies given by Searle. Thus, the consideration in turn of each possibility will automatically involve a rejection of both analogies.

3.1 Case 1: assume a necessary connection

What we are to assume here is that for some micro-level description of goings-on, X, and some description of macro-level goings-on, Y, it is a necessary truth that if X then Y. That is what I mean by there by being a necessary connection. This applies, for example, to the liquid case: it is necessary that a system of particles that have the properties that H_2O molecules have at standard temperature and pressure and behave as H_2O molecules do should be liquid.

Unfortunately, Searle is often able to get away with passing off the liquid analogy due to confusion on the part of audiences who do not understand the liquid case; so it is necessary that I explain the liquid case. What happens is that a conceptual analysis of a macro-level property is done; a model or theory of what could be going on at the micro-level is proposed; it is shown how the presence of the macro-level feature can be derived from that theory; and the theory is accepted as true, partly on the basis of its ability to explain the macro-level feature.

Thus, take the case of liquidity. There are three common states of matter; solid, liquid, and gas. A solid has a fixed size and shape and 'comes in one piece' - i.e., it is all connected together. It resists compression, shear, and penetration. If enough force is applied to it, though, it will deform permanently (get broken or bent).

A liquid has a fixed volume but not a fixed shape. It resists compression but not penetration. It has almost no tensile strength. If a shear is applied, it deforms immediately and continuously. It flows down and conforms to the shape of the bottom of its container, with a level surface at the top. Most liquids get you wet - i.e., a surface of liquid will stay on you if you touch it.

A gas has neither fixed shape nor fixed volume. It expands to fill its container. Under force it compresses or shears immediately and continuously and offers almost no resistence to movement into and through it.

Note that these descriptions are conceptual analyses of liqudity, solidity, and gaseousness, not experimental results:

I am saying these explain what it is to be liquid, solid, or gaseous, for any possible substance.

Next, a theory is proposed: Material substances are made up of a whole bunch of tiny particles called molecules, with big spaces between them, and these molecules have attractive and repulsive forces. The molecules are always vibrating with a certain amount of energy, this energy being a function of their speed and their mass.

Now the crucial point in all this is that a knowledge of what forces are on the molecules in what directions, what momenta they have, how they are spaced, their size, and how the forces between molecules vary with distance - knowledge of all that - is sufficient for predicting the properties of solidity, liquidity, or gaseousness, as defined above. Fluid dynamics is recognized as a special case of statistical mechanics. For example: The fact that the volume of a liquid is fixed can be explained by the attractive and repulsive forces among the molecules, specifically, the fact that the repulsive forces increase sharply with incremental decreases in the distances between molecules. However, the molecules do not have sufficient kinetic energy to escape from the group; they are held together by weak attractive forces. So the molecules will stay spaced at about the same distance, so that the whole substance will have a fixed volume.

The attractive forces are weak and do not increase with incremental increases in distance. This is why liquids do not come in one piece and have no tensile strength.

Explanations could be given for all the properties of liquids, solids and gases that I listed above. The theory explaining liquidity would consist in a description of the facts about the behavior of molecules in a liquid. And the reason why the theory is an explanation of liquidity is that given the theory about what goes on at the micro-level in a liquid, we can see why a substance would have to be liquid. For the same reason, liquidity can be said to be reduced to micro-level events in the strongest sense, namely, that of being a logical consequence of the micro-level events. So once the micro-level events have been described, nothing new is added by mentioning that the substance is liquid. So the liquidity is nothing over and above the sum of the actions of the micro-particles.

The necessary connection is one way: that is, the theory about the micro-elements necessarily implies the macro-level description ascribing liquidity (that is what makes the theory an explanation of the latter), but the observed liquidity does not necessarily imply the theory (that is why it is a theory instead of a logical deduction from observations).

Now that we understand the liquid case, we can see that it is utterly unlike the mind/brain case. There is no necessary connection between any physical events and any mental events, because for any physical description that could be given, it seems, it would be easy to conceive a possible world in which that description holds but no one is conscious. Neurons have only physical properties and no mental properties, and there is no way to derive a statement ascribing mental predicates from any set of purely physical statements. The way to see this is to just reflect for a moment on the physical concepts we have. To forestall irrelevant objections, I repeat that I am not claiming that there is no way that consciousness could be caused by physical events - that is not what I mean by "derived" - I am saying there is no way that consciousness could be logically deduced solely on the basis of physical events.

What does all this show? Well, first, the liquid analogy is entirely misleading, since one of the central, salient features of that case is lacking in the mind/brain case.

Second, if my claim that there is no necessary connection between the behavior of neurons and consciousness is true, then it follows that physical events cannot ever constitute an explanation of consciousness. If someone asks why people have consciousness, you can describe their brain chemistry, but the questioner will still be able to wonder, "Yes, but all that could have taken place without there being any consciousness, so why is there consciousness rather than just a mindless machine?" It is as if someone asked why the sky is blue and a logically unrelated fact were offered as 'explanation': "Because Paris is the capitol of France." Therefore, #5 in Searle's account is false. And given that consciousness has effects on our behavior, #6 must also be false. That is, the behavior of neurons doesn't explain consciousness, and since consciousness is essential to a correct explanation of our behavior, the behavior of neurons does not explain our behavior.

Third, supposing that we just throw out the analogy and subtract claims (5) and (6) from Searle's account - that is, if we take the false items out of his account - what we are left with is a set of fairly obvious and unilluminating claims. We are left with an account that says brain activity causes mental events, brains have consciousness, neurons don't, and mental states are irreducible. This is hardly a solution to the mind/body problem. We are still left with Descartes' problem - the one that can be fairly said to have started the entire philosophic debate - namely, how can these purely physical states interact causally with states in the distinct, non-physical, mental realm?

Searle's reply

Searle responds to essentially the objection I have just voiced in chapter 4, section III of the reader. Let's consider his replies in turn.

(1) He briefly disputes the fact that scientific explanation requires necessity, pointing out that there is no necessity to the Law of Gravitation, for instance. However, this example seems rather to support my case than to refute it. My claim was that in order that some theory (or any statement) should explain something, the connection must be necessary: i.e., it must be necessary that if the theory is true then the situation explained exists. I did not say that the theory must itself be a necessary truth. Now the Law of Gravity may not be a necessary truth. However, it does explain the motions of the planets and the way things fall to the ground because the law necessarily implies those phenomena.

(2) He suggests that "the apparent 'necessity' of any scientific explanation may be just a function of the fact that we find the explanation so convincing that we cannot, e.g., conceive of the molecules moving in a particular way and the H_2O not being liquid." I don't understand what he is talking about. We don't regard the connection as necessary because we believe the explanation. We only accept the explanation because of its ability to explain, which is due to the necessary connection. Searle suggests that a person in antiquity might not have found the explanation a matter of 'necessity'. All I can say is that if he didn't, he'd be wrong. I understand modalities to be objective facts about propositions. It isn't a psychological claim. The question to ask is, "Could this have been otherwise?" Now, if some other person would answer yes to that question where I answer no, then I say he is simply wrong. But I can't see why we should expect that a person in antiquity would have been unable to perceive the necessity. I suggest that Professor Searle is simply confused about what necessity is. His remarks lower down confirm this suggestion. He writes:

> [I]f I see a screaming man with his foot caught in a punch press, then I know the man must be in terrible pain. It is, in a sense, inconceivable to me that a normal human being should be in such a situation and not feel a terrible pain. The physical causes necessitate the pain. (p86)

But this argument confuses physical or nomological necessity (truth in all possible worlds in which the laws of physics hold) with absolute necessity (truth in all possible worlds). Of course, the physical causes cause the pain, but they do not logically necessitate it. And it is just obviously false that I cannot conceive of someone being in such a situation and not feeling pain. I can conceive of it easily, though I would not believe it. It is telling that Searle adds "in a sense" and "normal". Depending on how "normal" is understood, perhaps it is necessary that a normal human being would be in pain, but it is not necessary that that human being be in pain since it is not necessary that he be normal.

(3) Searle suggests granting the point about the lack of necessary connection for the sake of argument. Then he goes on to say, "Nothing follows about how the world works in fact. The limitation that Nagel points out is only a limitation of our powers of conception." This suggests that he has a subjectivist or psychological theory of necessity - i.e., that he thinks something's being necessary is solely a matter of what we can or can't conceive. This is not the case. It is not a matter of whether we can see something must be true; it is a matter of whether it must be true. It is a matter of whether something actually could be otherwise. The fact that we can conceive of something is not what possibility consists in; it is only a way of finding out whether something is possible. After all, I might be able to imagine the four-color theorem being false, but all the same, given that it is true, it is necessarily true. We could be wrong about whether something is necessary, of course, just as we can be wrong about anything. But I am arguing that it seems as if the mind/brain connection is not necessary. The two replies that are suggested by this cryptic passage of Searle's are (1) that necessity or lack of necessity is no fact about the world but only about how we think, so it doesn't say anything about the nature of the mind and (2) that our apprehensions of necessity or lack of necessity are illusory. It is obscure which of these, if either, Searle is saying. But as for the first, I simply don't agree. I think whether something could have been otherwise is just an objective fact about it, not about how we think. As for the second, of course, it is possible to claim that any intuition or perception or other cognition is an illusion, but the burden of proof is on who makes that claim.

Searle seems to suggest at one point that there really is a necessary connection but that the reason we cannot see this is that we are inside consciousness, and in order to 'picture' the necessary connection we would have to get outside consciousness and picture both the mental and the physical facts from outside. This doesn't make any sense to me. I don't see in the first place why Searle thinks there is a necessary connection. He doesn't give any argument for it, and intuitively it seems there is none. In the second place, I don't see how the fact that we are conscious means that we

could not see the necessary connection if there were one. Granted, we cannot get outside of our minds, but so what? We can still understand the nature of our minds. This is shown by the fact that we actually do have mental concepts that we understand. And we can still understand the nature of the physical world. So we can understand both terms of the relation. And we seem in general to be able to distinguish necessary truths from contingent ones. So why should we not be able to see that the fact that brain processes cause consciousness is necessary, unless because it is not necessary?

Searle asks us to imagine a machine that could detect causally necessary relations. He says that it would see no difference between matter/matter forms of necessity and matter/ mind forms of necessity. But we are not interested in 'causally necessary relations' if that just means "causal relations"; what we are interested in is necessary causal relations. Given that, the machine would detect a difference: it would detect that matter/matter forms of necessity exist, while there are no matter/ mind forms of necessity.

Before we get carried away with this point, though, we must remember that I have been talking as if all cases in which an observable feature is caused by an underlying microstructure exhibited a necessary connection, whereas this is not obviously the case, and Searle has given an analogy in which there is seems to be no necessary connection. We must now consider this type of case.

3.2 Case 2: assume no necessary connection

What we are to imagine this time is that for some micro-level description of goings on, X, and some macro-level description of events, Y, it is contingently true that if X then Y. This situation is exemplified by the heat case (or at least, so it seems at first glance): heat is the mean kinetic energy of molecules, but one cannot logically deduce that something is hot solely on the basis of knowledge of mean molecular kinetic energies. Heat could have turned out not to be motion of molecules.

There are a few different ways of conceiving of this case corresponding to different theories about primary and secondary qualities and so on, but whatever plausible theory of what is going on in the heat case is adopted, it is not analagous to the mind/brain case. Let's look at the possible interpretations:

(1) Kripke argues in *Naming and Necessity* that heat is necessarily identical with molecular motion, though this fact is known a posteriori. On this model, the mind/brain case is disanalagous because, as Kripke also argues and as I have argued above, it is not necessary that brain processes correlate with mental states.

(2) Next, there is the theory which says heat is just a sensation. On this account, heat is therefore not motion of molecules. There will be no necessary connection between heat and molecular motion on this view, and therefore, according to my account of explanation, there is no explanation of heat. However, notice that this just makes the irreducibility of heat a special case of the irreducibility of consciousness: heat is a sensation and, just like all other mental states, it cannot be explained by physical events. Because heat is a mental event, it is absolutely useless to be told, "You want to know the relation of mental phenomena to physical phenomena? Well, it's just like the relation of heat to physical phenomena." Moreover, the analogy should actually read

consciousness : brain processes :: heat : brain process x

That is, whatever brain process causes the sensation of heat, that process stands in the same relation to heat as brain processes in general stand to consciousness in general. Mean molecular kinetic energy is the wrong analog. If heat is a sensation, the relation between heat and motion of molecules is that the former is used to detect the latter, whereas consciousness is not used to in that way detect brain processes, so the analogy is false.

(3) Next, there is the view (held by Locke, for instance) that heat is the power to produce a certain sensation (the sensation of heat) in human minds, and it turns out that molecular kinetic energy is what gives objects this power. If this were analagous to the relation between consciousness and brain processes, then we could say that consciousness is just the power to produce the sensation of consciousness in human minds, and brain processes have this power. But that doesn't make any sense.

(4) Searle's own view says that initially, "heat" referred to a sensation. We found out that it was the motion of molecules that caused this sensation. Then we redefined "heat" to refer to the motion of molecules. On this view, to avoid equivocation, Searle must still decide what he wishes to use "heat" to refer to when he makes the analogy. He may conform his usage to the past usage or to the present usage of the word, but either way, he will have to face one of the arguments I have just given above. Either heat is a sensation, in which case the analogy is circular as an explanation besides picking the wrong analog; or it is motion of molecules, in which case it necessarily implies motion of molecules.

Searle presents a slightly different picture from either of these alternatives, though. The picture he paints seems to say that heat has both a subjective, phenomenal aspect (its surface appearance) and an objective aspect (motion of molecules). We don't find the subjective aspect very important or interesting, so when we find out about the objective part, we are inclined to kick the subjective part out of our conception and say heat is just motion of molecules. In an exactly parallel way, says Searle, pain has both a subjective, phenomenological aspect (what it feels like) and an objective aspect (brain processes). However, in this case we do find the subjective aspect interesting, so we are not willing to eliminate it and say pain is just a brain process. This is why pain can't be reduced to brain processes while heat can be reduced to molecular motion -- but it does not mean that the cases are disanalagous with respect to any intrinsic, metaphysical facts; it just means that we take different attitudes and apply different definitional practices.

I have tried to do justice to this uncharacteristically confusing and confused passage. The confusion I think Searle is committing can be illustrated by comparing the use of "sensation of" in the following two expressions:

a sensation of red	a sensation of pain

A sensation of red is a sensation that is of red, i.e., it senses red. The sensation, of course, is not itself red. A sensation of pain, however, does not sense pain; rather the sensation is the pain. Similarly, compare

an experience of a yellow station wagon	an experience of euphoria

An experience of a yellow station wagon is not a yellow station wagon, but an experience of euphoria is euphoria. On the one hand, "of" is used to denote intentionality; on the other, it denotes identity. My sense is that Professor Searle has confused these two relations vis-a-vis

the subjectivity of heat	the subjectivity of consciousness

Of course, the subjectivity of heat is not heat, nor is it something that inheres in heat; rather, it is a facet of our experience of sensing heat; whereas the subjectivity of consciousness is consciousness.

Searle's explanation of why heat really is analagous to pain depends on both of them having a subjective aspect. But to say they both have subjective aspects is an equivocation. "The subjectivity of consciousnss" means what it's like to be conscious. "The subjectivity of heat" does not mean what it's like to be hot; it means what it's like to touch (detect, encounter, etc.) something that is hot. Now we can name the metaphysical disanalogy between the cases of heat and consciousness. Consciousness has subjectivity as an intrinsic, real feature of it. In the relevant sense, there is no subjectivity in heat - i.e., there is no such thing as what it's like to be hot in the sense in which there is something it is like to be conscious. The phenomenal aspect of heat is not an intrinsic feature of motion of molecules, but consciousness is an intrinsic feature of the brain. The reason heat can be reduced is that ontologically, it doesn't actually have any subjectivity. But the mind does.

(5) Suppose that heat is conceived as a real, objective property of things (not a sensation) that is contingently identical with mean molecular kinetic energy. Then the relationship is that "mean molecular kinetic energy" and "heat" are two descriptions of the same phenomenon, but the former is the one that identifies the essential, underlying reality, while the latter identifies the phenomenon by appearances. If the mind/brain case were analagous, we would have to say that brain processes were the reality underlying the appearances of consciousness. But this makes no sense, not only because the appearance/reality distinction presupposes the concept of consciousness but also because the appearances of consciousness are necessarily identical with the real nature of consciousness. It does not make sense to try to find out what consciousness is really like as opposed to what it seems like to us, because the way it seems is its ultimate reality.

Summary

We'd better take a retrospective look now at the ground we have been over. Searle compares the mind/brain relation to the liquid/H_2O relation and the heat/molecular motion relation. We have seen why these comparisons are inaccurate. He also claims that a scientific explanation of consciousness in terms of the behavior of neurons can be given. We have seen that this is impossible because there is no necessary implication from any possible behavior of neurons to any conscious states. This leaves us stuck with the old mind/body problem, as intractable as ever. Searle makes two attempts to attribute the mysterious features of the mind to epistemic problems rather than inherent facts about the mind's unique, strange nature. The first of these attempts is the claim that there really is a necessary connection between mental and neurological events but we can't see it because we cannot get outside our consciousness in order to picture it. I argued that this account is unsatisfactory and depends either on assuming that the existence of necessity is some psychological fact or on assuming that our perception of necessity is illusory. The first is a misunderstanding of the concept of necessity, and no argument is provided for the second. The claim also fails because the fact that we cannot get outside something is no impediment to our understanding it or to seeing its necessary relations to other things.

The second attempt to attribute the apparent mysterious features of the mind to purely epistemic problems is the claim that the irreducibility of consciousness is a consequence of our definitional practices. This claim fails because it depends on a confusion about subjectivity that leads to a failure to recognize that consciousness is the only thing in the world with subjectivity as part of its inherent nature.

So the analogies fail, the interesting part of the description of the relation of mind to brain is false, and the attempts to explain away the mysteries that give rise to the mind/body problem fail.

September 20, 1992

Dear Bryan,

I guess the enclosed article confirms my theory (but see later remarks for help in evaluating this claim), don't you think? It's really perfect; there is nothing I can add.

By the way, what is this "marginal cost pricing/average cost pricing" business?

I appreciated the conclusion of your paper, though I expected you to also make some other traditional points about the problem of people having bad values - e.g., the fact that, even on the assumption that people's values are bad, there is no other species that can be put in charge; and also that, notwithstanding fallibility, there is always a tendency for beliefs (and values) to be true since, presumably, people don't just pick beliefs at random. Hence, it is highly improbable that people would not for the most part prefer good things to bad (the unusualness of your Lecter example illustrates this ... though I, of course, would rather turn the brats over to the doctor, even if he didn't pay).

You could stand to fill in your argument a little more detailedly (though perhaps that would take it away from the field of economics). For instance, you probably know of a few suitable historical examples, including the cases you were reading about of successful non-violent social movements (indicating that people's opinions were effected not by force but by examples of action from principle). Religions in some cases illustrate the idea also: Christ's example, like the example of an honest, benevolent, good priest (as in *Les Miserables*) exerts a much more profound attractive influence on people toward religion than censorship ever did or, I would guess, the Crusades. Finally, you need to make out clearly the mutual incompatibility of the two approaches, of the argument of force and the force of argument. It is hard to listen to someone and learn to respect them when you're fighting with them, hiding from them, and/or unable to freely discuss your views.

Someone might, by the way, object by asking rhetorically whether you would suggest that, as your argument would appear to imply, we try persuading the unregenerate cannibal/serial child murderer to convert to our ways by merely conspicuously refraining from eating children. This is not apt to work. To which you might reply that in some sense, your argument was meant to apply to the realm of serious disagreements among reasonable people, not to the case of sheer psychosis or radical evil.

* * *

I was planning on sending you a letter about confirmation theory, which I am just learning about now. I'm reading *Scientific Reasoning: the Bayesian Approach* by Colin Howson and Peter Urbach. I'd like to explain it to someone else because I like it so well.

Confirmation theory is supposed to be analagous to formal logic. The latter investigates the relation of entailment and tries to state under what conditions it exists between propositions. The former similarly investigates a weaker relationship between propositions, viz. confirmation, which exists between premises and conclusions of valid inductive arguments.

Unfortunately, confirmation theory has had some difficulties, as you might have expected. Here is one of them:

The Ravens Paradox:

Nicod's criterion, named after Jean Nicod, lays down the following plausible general rule(s):

(a) The observation of an A that is B confirms "All A's are B."
(b) The observation of an A that is not B disconfirms that all A's are B.
(c) The observation of a non-A is irrelevant to, i.e., neither confirms nor disconfirms, that all A's are B.

These seemingly trifling truisms have the following implications, in particular:

(1) The observation of a black raven confirms that all ravens are black (by condition (a)).

(2) The observation of, say, a white chair does not confirm that all ravens are black (by (c)).

However,

(3) The observation of a white chair (which is neither black nor a raven) does confirm that all non-black items are non-ravens (according to (a)).

Yet "All ravens are black" is exactly equivalent to "Everything that isn't black isn't a raven." So the same observation, according to Nicod's criterion, confirms a proposition even though it fails to confirm a logically equivalent proposition. (Incidentally, this also of course violates the weaker and at least equally obvious rule that an observation that confirms some proposition should also confirm the logical consequences of that proposition.) This is found to be paradoxical, particularly given the plausible general rule which we can call the Equivalence Condition:

(d) If e confirms h, then e confirms everything that is logically equivalent to h.

So one of these plausible rules ((a) - (d)) is false. What do you think? I encourage you to think it over before I give you the right answer, which I will give in the next paragraph, though without explanation or defense. Carl Hempel thinks it's (c) - i.e., he thinks that the observation of a white chair does confirm that everything that isn't black isn't a raven and also, consequently, that every raven is black. This leads to indoor ornithology.

I think the problem is the unqualified statement of (a); i.e., it depends what A and B are, as to whether the observation of an A that is B confirms that all A's are B.

I'll let you think this over for a while.

Here's another problem of confirmation theory:
Carl Hempel discusses the following initially plausible rules:

(1) The consequence condition: if evidence confirms a proposition, then it confirms the logical consequences of that proposition.

(2) The prediction condition: if evidence confirms a consequence of some hypothesis, then it confirms the hypothesis.

These two principles, however, have the unfortunate consequence that everything confirms everything. For suppose we discover some proposition, e (for "evidence"), to be true. Take any hypothesis, h. By the prediction condition, e confirms (e & h) since e follows from (e & h). And since h follows from (e & h), by the consequence condition e also confirms h. This is odd since we haven't said anything about the content of e or h.

Here I think the problem has to be the prediction condition. Taken without qualification, that must be false. (Again, it depends on what the hypothesis and the prediction are, as to whether the truth of the prediction confirms the hypothesis.)

I haven't said anything about the Bayesian theory yet. The Bayesians are considerably more informative than Nicod or Hempel. They base their entire theory of how scientific and all inductive inference works on a certain result of probability theory, Bayes' Theorem, which I can explain to you fairly easily.

The 'probability calculus' has four axioms. Letting P(x) stand for the probability that x is true and P(x | a) stand for the probability of x given a, i.e., the probability, if a is true, that x will also be true, the axioms are

(1) The probability of anything is greater than or equal to zero:

$P(a) > 0$, for all a.

(2) The probability of any necessary truth is one:

$P(n) = 1$, when n is necessarily true.

(3) If a and b are mutually exclusive alternatives, then the probability that one of them is true is the sum of their individual probabilities:

$P(a \text{ or } b) = P(a) + P(b)$, if a and b are mutually exclusive.

(4) The probability of a and b both being true is the probability of a being true times the probability of b given a.

$P(a \ \& \ b) = P(a) * P(b \mid a)$.

Hopefully, you can see these facts intuitively. Incidentally, there is a series of clever arguments ("Dutch Book arguments") that try to show that any rational person would choose to have his degrees of belief correspond to the probability calculus. They show that if you accept bets at odds determined by your degrees of belief (e.g., if you think it 90% likely that it will rain and 10% likely it will not, you are willing to take either side of a bet on its raining at 9 to 1 odds), which is in accordance with Bayesian decision theory of course, then any time your degrees of belief violate one of the above axioms it will be possible for someone to offer you a bet which you would be willing to accept and in which you were guaranteed to lose money. However, it's not necessary to go into the details of that argument, since the axioms are indubitable anyway.

Bayes' theorem is a simple result of (1) - (4): take two propositions, e and h (for "evidence"

and "hypothesis").

P(e & h) = P(h & e), obviously. Invoking (4), we can substitute on both the left and right sides to obtain:

P(e) * P(h | e) = P(h) * P(e | h). And rearranging, we get

$$P(h \mid e) = \frac{P(h) * P(e \mid h)}{P(e)}$$

which is Bayes' Theorem. The Bayesians interpret this result as follows. Suppose we've just discovered some evidence, e, and we want to (re-)evaluate our hypothesis, h, in the light of this new evidence. Then the posterior probability of h, given e, is directly proportional to the prior probability of the hypothesis, and to the liklihood of the evidence on the assumption that the hypothesis were true, and inversely proportional to the initial probability of the evidence (without the assumption of the truth of the hypothesis). Thus, for confirming our theories, we want an initially plausible hypothesis and we want evidence that is highly likely if the theory is true but unlikely otherwise.

Unfortunately, the Bayesians, at least the ones of the school to which the authors of the book I am reading belong, do not provide any general method for the assignment of prior probabilities, and, worse, they don't believe that any such general, uniquely rational method exists. All they can do is tell you how to update your degrees of belief, but not how you should start out. It seems that their theory is that you can start out in life with any initial probability distribution you want. And although there is a tendency for large amounts of evidence to wash out the significance of the priors, there still are sets of prior degrees of belief which are such that virtually any evidence could accord with virtually anything, given those priors, or virtually anything could fail to be confirmed. For instance, obviously, if you start out thinking that the probability of any universal generalization is zero, then no amount of evidence can ever change your opinion. (Plug in 0 for P(h) in the above statement of Bayes' theorem, and P(h | e) is always 0.) This would be the case, for example, when it comes to tosses of a coin: no matter how many times the coin has come up heads, the probability of the next toss coming up heads is unchanged. And the probability that it ALWAYS comes up heads for the indefinite future is zero (1/2 raised to the power of infinity), and stays zero no matter what. Thus, it would appear that no scientific law is rationally demonstrable - i.e., a rational being could continue to doubt it no matter what. (Though according to the Bayesians, another rational being could accept it, given the same evidence.)

I haven't yet gotten to the part where they answer this objection, but they promise to argue that their theory is not "too subjective".

By the way, it appears evident that an element of intuitive judgement will be necessary to acquire degrees of belief prior to acquiring evidence. Perhaps I will provide more discussion of this some other time.

Wisely,