When Is a Brain Like the Planet?*

Clark Glymour†‡

Time series of macroscopic quantities that are aggregates of microscopic quantities, with unknown one-many relations between macroscopic and microscopic states, are common in applied sciences, from economics to climate studies. When such time series of macroscopic quantities are claimed to be causal, the causal relations postulated are representable by a directed acyclic graph and associated probability distribution—sometimes called a dynamical Bayes net. Causal interpretations of such series imply claims that hypothetical manipulations of macroscopic variables have unambiguous effects on variables "downstream" in the graph, and such macroscopic variables may be predictably produced or altered even while particular microstates are not. This paper argues that such causal time series of macroscopic aggregates of microscopic processes are the appropriate model for mental causation.

1. Can There Be Mental Causes? All of us talk as if some thoughts cause some actions. We distinguish deliberations that guide a course of action from random thoughts, fantasies, rejected plans, and even intended consequences that are brought about by our intentions but in ways not intended. We say that the causal role of some of our thoughts is part of

*Received February 2005; revised May 2007.

†To contact the author, please write to: Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: cg09@andrew.cmu.edu.

‡I am grateful to Jim Woodward, Michael Strevens, Elliott Sober, and John Campbell for valuable comments and corrections to a previous draft of this essay and to the participants in a conference on fMRI issues sponsored by the Rutgers Department of Philosophy in 2006, and especially to Steve Hanson, for stimulating discussions that caused substantial revisions in this essay. The graph of Figure 5 was obtained by Joseph Ramsey by applying a Bayesian search procedure, the Greedy Equivalence Search algorithm, implemented in the TETRAD IV program (http://www.phil.cmu.edu/projects/tetrad) to the lagged variables of unpublished data from Steve Hanson's laboratory. The title of this essay was offered as a joke by Mara Harrell, but I took it seriously. Research for this paper was supported in part by a grant to the University of California, Berkeley from the James S. McDonnell Foundation, by grants from the National Aeronautics and Space Administration to Carnegie Mellon University and to the Florida Institute for Human and Machine Cognition, and by a grant to Rutgers and Carnegie Mellon from the James S. McDonnell Foundation.

Philosophy of Science, 74 (July 2007): 330–347. 0031-8248/2007/7403-0001$10.00

their very content, as when one has the thought of trying to do something. Judgments about mental causes—motives—are woven into systems of law and informal customs of praise and blame.

Times change, and with them accounts of whether and how reasons can be causes. A century ago an eloquent claim to a vital force, evidenced by the mind, with causal powers well beyond those of conventional physics, was worth a Nobel Prize—at least in literature.[1] Nowadays, Templeton prizes, not Nobels, are quaintly given for vitalist projects; scientific Nobels are given for chemical explanations of how aspects of mind come about. Against the common sense that thoughts are sometimes causes, contemporary psychologists describe a variety of experiments showing that actions can be caused by something other than conscious thoughts.[2] Neuropsychologists add further considerations. In experiments measuring brain activity during simple judgment tasks, conscious awareness is anticipated by characteristic neural events (Libet 2004), and in experiments presenting participants with a narrow set of alternatives, the content of perceptual judgments can be predicted from magnetic resonance images of the brain (Suppes et al. 1997, 1998, 1999; Suppes and Han 2000). And, finally, late-twentieth-century philosophy has generated arguments against the very possibility that mental properties can be causal factors. The question is then in what sense, if any, the occurrence of the properties we call mental can be causes of anything. I will attempt an answer, which in summary is this: Property identifications are local, not universal; locally, occurrences of mental properties are aggregates of occurrences of neural properties; aggregates can have causal relations that none of their constituents have, and mental properties do so. I claim that the form of the answer conforms pretty exactly to causal claims in everyday science apart from neuroscience, and the substance of the answer conforms equally well to the leading edge of current neuropsychological explanations.

2. Local Identifications and the Philosophical Argument. In the recent philosophical literature about—against, really—mental causation, there is a kind of skeptical master argument that goes something like this:

1. Actions are (at least) physical events.
2. The joint occurrences of physical properties of physical events are always sufficient causes of physical effects.

1. See Bergson ([1911] 1998). See also the work of the same period by the great South African statesman Jan Smuts (1973).

2. The famous source is Michael Faraday's demonstration that séance sitters mistakenly judged the forces they applied to rapping tables, a kind of argument once rare; nowadays related demonstrations are routine (Wegner 2003).

3. If, for every sufficient set of physical causes of a particular event, there is a set of physical events that are sufficient causes for each member of the first set, only physical events are causes of the particular event.
4. Two properties are identical if and only if they are necessarily identical.
5. No mental property is necessarily identical with any combination of physical properties.
   Subargument:
   5.1. We can imagine any mental property to be realized in physically different constituents than brains.
   5.2. Whatever is imaginable is possible.
   5.3. Therefore, no mental property is necessarily identical with any combination of physical properties.
6. Therefore, no mental property is identical with any combination of physical properties.
7. Therefore, joint instances of mental properties are not causes of action.

An addendum asserts the anomalism of the mental: there are neither deterministic nor statistical psychophysical laws that reduce any mental property to physical properties.

The master argument has any number of variations.[3] The subargument

3. The master argument, and the functionalist and second-order property responses, are cobbled from many, many essays in the philosophy of mind, including many of those collected in Heil (2004) and Chalmers (2002) as well as book-length essays on the question by Kim (1993, 1998, 2005) and of course Kripke (1980). The thesis of the anomalism of the mental is due chiefly to the influence of Donald Davidson, but the essay (1970) in which the doctrine is announced contains no argument I can find other than the absence of any such established laws. Davidson requires that laws be "strictly universal" and treats it as more or less obvious that the required universal equivalences do not exist. Published in 1970, the essay reflects the common view of the time that the relevant laws, if any, would be behaviorist on the physical side, not neuorophysiological, and its influence may be due to saying vividly what philosophers of the time already believed. The issues that arise in considering the master argument will encompass Davidson's assumptions.

The master argument has at least one important variant for (5), essentially due to Thomas Nagel (1974):

I. If all truths about one class of properties could be known without knowing truths about any of a second class of properties, no property in the second class is a member of the first or reducible to members of the first.
II. Knowledge of all truths about the physics, chemistry, and neurophysiology of a sentient being would not suffice for knowledge about the conscious mental properties of the being—for knowledge of how it feels to be that being.

can be defeated by denying that systems of other physical constitution or structure can have our mental states and properties, or by denying that what is conceivable is therefore possible. I endorse the second objection, but it does not go to the heart of the matter: were there aliens or robots physically different from humans but sharing human mental states, the identity of mental properties with physical properties would not be disproved, because property identity is local, not global.

The temperature of a gas is the mean kinetic energy of the molecules of the gas. In gases, temperature and mean kinetic energy are the same property—but not in radiation. Radiation has a temperature, but the temperature of radiation is not the mean kinetic energy of the radiation. Temperature is a quantity that may be measured in myriad ways, with different connections to other quantities in ways we cannot delimit, defying a disjunctive definition. Temperature is not identical to mean kinetic energy or to frequency of radiation, and so forth. Rather, the temperature of a gas at equilibrium is the mean kinetic energy of its gas molecules. Light *is* electromagnetic radiation, but the identity is not global: not all electromagnetic radiation is light. Sound in the atmosphere is identically the vibration of the molecules of air, but sound in water is no such thing. Nor is sound just any vibration: atoms in crystal lattices vibrate soundlessly, although their vibrations can in some circumstances cause acoustic vibrations. What we regard for good reasons as different instances of the same property can in one instance be identical with another property and in another instance not. These local identifications are not like the penniful property of being a coin in my pocket and being made of copper. Property identifications are conditional, but in the conditions they are necessary, not contingent. The identity of sound in air and the vibration of air molecules, light, and electromagnetic radiation cannot be made otherwise. Instances of the property can be destroyed (eliminate the air), but do what you will, as long as you have vibrating air you have sound. It follows that it is at least conceivable that one and the same mental property could be identical with different physical properties in humans and in aliens or in robots; indeed, one and the same mental property could be identical with different physical properties in you and in me, or in any one person at different times.

III. Therefore, no conscious mental property is identical with any combination of physical properties.

The "gap" is sometimes described as "explanatory," but it is really existential. The gap between consciously knowing a physical description and what is known by being what is so described is not likely to be filled by anything, however well we come to understand the physiological conditions that produce consciousness in terrestrial species. The variant argument applies only to conscious mental states and processes. Accordingly, I focus on the master argument, but I feel the gap.

If the explanation of mental phenomena by cognitive neuroscience is possible and if mental events are causes and their mental features have causal roles, there must then be some criteria for the local, conditional identity of mental and physical properties, and for such identifications to be discoverable there must be enough stability to the identities of mental and physical properties so that evidence can be acquired that the criteria are met. Not everything going on in the brain is mental; all sorts of physiological properties, events, and processes are correlated with mental phenomena but should not be identified with any. All sorts of mental events appear to have no influence on action, and conceivably, all sorts of mental properties have no causal role. Criteria for sorting seem wanted.

In an essay elaborating why it is that the fact that one can imagine that two properties are not identical does not imply that they are distinct, Ned Block and Robert Stalnaker (1999, 29) claim that identities between conscious mental states and physical states might be justified by "the same kinds of considerations that are used to justify water = H_20." (Water is only locally identical with H_20, of course, but never mind for the moment.) The considerations they refer to are vaguely characterized as "simplicity" and "best explanation." Jaegwon Kim (2005, 142) waxes almost irate at the suggestion: "This proposal is bold and surprising—and more than a little incredible! . . . [I]t is difficult to believe that a problem that has long vexed so many great minds in western philosophy, including some of the finest scientists, dividing them into a host of warring camps, should turn out to be something that could have been solved the same way that scientists determined the molecular structure of water." Nothing makes some philosophers less happy than the prospect that a philosophical problem might actually be solved. Granting that they cannot be disproved a priori, Kim cannot find in Block and Stalnaker's essay, or apparently by his lights anywhere else (Hill 1991; McLaughlin 2001), an account of explanation that would "justify" such identifications. Appeals to "simplicity" and "best explanation" are so much pen waving, he seems to think, and that far I agree with him. I suggest that scientific practice contains a principled scheme—more precise than "simplicity" and "best explanation"—for the identification of properties and their assignment of causal roles, and that mental causation plausibly falls within its scope.[4]

3. Causal Explanation, Not "Intertheoretic Reduction." One view about the relation between neuroscience and "folk psychology"—the wealth of

4. Kim's discussions of mental causation avoid all scientific details, but if an explicit replacement for the place-holding "simplicity" and "best explanation" is demanded, the demander is, I think, obliged to consider the statistical and scientific details that might fill the places.

everyday attributions of beliefs and desires and motives with which we explain our own and others' behavior—is that neuroscience aims at a "theoretical reduction," something like the relation between statistical mechanics and classical thermodynamics, or special relativistic kinematics and Newtonian kinematics.[5] Philosophical accounts of intertheoretic reduction from the 1960s and 1970s supposed two theories and some semiformal relation between them: one theory supplemented by "bridge laws" or other correspondences would entail the other, or would entail the other as a limiting case, or would provide formal "analogues" of the claims of the other, or would specify relational structures that could be mapped onto relational structures specified by the other. Bickle (1998) appropriates the analogy story to specify the relation between mental properties and physical properties, and Block and Stalnaker come close to doing the same, to which Kim objects that all the "explaining" is in the language of the reducing theory, and the regularities of mental phenomena remain unexplained.

For several reasons, it is a mistake to try to use these traditional logical schemes to frame the structure of what would be required for neuroscientific explanations of mental contents and their causal roles. "Folk psychology" is entirely unlike a scientific theory. On the one hand, folk truths are too general and banal—as with "people use their beliefs to try to obtain what they desire"—and on the other hand psychological truths can be too idiosyncratic—"madeleines bring back a flood of remembrances of things past." The robust generalizations of human and animal psychology are neither banal nor idiosyncratic, and often they are not what people believe about themselves and about one another; they are outside of folk psychology. Further, unlike, say, the reduction of Newtonian kinematics to special relativistic kinematics, the explanations that neuroscience aims to provide for mental life are causal; the goal is to describe the actual mechanisms of thought and to identify processes of thought of various kinds with the functioning of such mechanisms. Causal explanations have a special structure and a special methodology; they are not a matter of exhibiting one equation as an analogue or limiting case of another. While there may be relevant physical analogues—I will suggest one shortly—the connections we should look for between the mental and

5. The most extended recent presentation of this view is Bickle (1998). An essential part of Bickle's view is that the "reduction" makes no reference to the distinct language of the reduced theory; the explanation consists entirely in demonstrations within the language of the reducing theory. Separately, analogies between the results of the theories are noted. Essentially the same idea but in more elaborate logical clothing was presented by Ned Block and Robert Stalnaker (1999). Bickle's book is not cited, perhaps because it appeared too late; Kim (2005) devotes much of a chapter to the idea, citing only Block and Stalnaker.

the biochemical and neurophysiological will not be limiting case derivations of equations; nor will they be illuminated by algebraic manipulations on relational structures for the language of neuroscience and the language of mind. They will be causal explanations that display the pieces and processes through which kinds of thought come about and are constituted. Eric Kandel, the doyen of the biochemical study of learning and memory, said about the same (Kandel and Hawkins 1992, 79): "The biological analysis of learning and memory requires the demonstration of a causal relation between molecular mechanisms in neurons of the brain implicated in a particular form of learning and the modification of behavior produced by the learning." Kandel spent much of his career identifying the neural and biochemical mechanisms of flexible behavior—reasonably called learning and memory—in the sea slug, *Aphlysia*, focusing on mechanisms that cause the siphon of the animal to withdraw under its mantle—the sea slug equivalent of ducking. Hawkins and Kandel (1984) argued that various hypothetical cascades of cellular facilitation and inhibition of release of neural transmitters—the chemical mechanisms of which are generally understood—could account for a range of phenomena known for classical conditioning, including secondary conditioning and blocking.[6]

This work has been cited (Bickle 1998) as an example of "intertheoretic reduction," when to all appearance it is a straightforward proposal of a scheme for causal explanations, much as having shown that a simple clock works by springs and gears; one might speculate about how springs and gears could be put together to make a clock that shows the date as well as the time or to make a clock that shows the time in multiple time zones, and so forth. The causal part is in the mechanisms and submechanisms and their relationships; the assembled mechanisms, working normally, may *be* a clock.

James Woodward (2003) has argued that intervention relations are necessary conditions for causal relations; A causes B only if B varies with some possible intervention on A. I disagree, but I think that intervention relations are bound up with both necessary and sufficient conditions for property identity. Beyond some stable correlation,[7] property identification

6. Secondary conditioning occurs when a kind of event (secondary stimulus) associated with a kind of event (primary stimulus) with which a pleasant or unpleasant kind of event (unconditioned stimulus) is associated itself becomes associated with the unconditioned stimulus. Blocking is the following phenomenon: once occurrence of a property has become associated with an unconditioned stimulus, co-occurrence of that property with another property followed by the unconditioned stimulus does not result in learning an association between the second property and the unconditioned stimulus.

7. Hill (1991) and McLaughlin (2001) have each argued that identity is the best explanation of strong correlations of properties. I require more.

requires correspondence of effects under hypothetical or actual manipulation: If, under conditions *C*, *A* causes *D*, then if under those conditions *A* and *B* are the same property, under those conditions manipulations of *A* that alter *D* should correspond to manipulations of *B* that also cause *D*, and vice versa. Further, for identity of mental properties or processes with aggregates of physical properties or processes, the time order and statistical relations of the occurrences of mental properties or processes must be the time order and statistical relations of the aggregates of the physical properties or processes.

The first comes about as follows. Properties can be strongly correlated without being identical. The length of a flagpole's shadow is strongly correlated with the height of the flagpole and the altitude of the sun, but not identical with any complex or function of either. If the height of the flagpole is changed (telescoping flagpole!) or the height of the sun changes, the length of the shadow changes. But the length of the shadow can readily be changed without any change in height of the flagpole or the sun (introduce an angled surface on which the shadow falls). The asymmetry is a mark—indeed a sufficient condition—for the shadow length not to be identical with any property determined by the flagpole height and the sun altitude. The lack of such an asymmetry is not, however, a sufficient condition for property identity. Richard Scheines and Peter Spirtes (2002) have offered the following example, taken from the state of medical science some years ago. Suppose that a medical researcher advanced the hypothesis that elevated cholesterol levels cause heart attacks. Several drugs are known to lower the total cholesterol level in the body and to have no other direct effects on heart attack rates. In an experiment, total cholesterol blood levels are measured in randomly selected subjects: some of them have been given recommended dosages of several drugs, whereas some have been given only placebos. The subjects are then followed over time and the rate of heart attacks in the various groups is calculated. Suppose it turns out that different drugs have different associations with heart attack rates, and, overall, heart attack rates are not independent of the treatment conditional on the resultant cholesterol level. What can be the explanation?

Suppose in fact that low-density cholesterol causes heart attacks but high-density cholesterol has no effect. The actual causal structure in the experiment is seen in Figure 1.

Various drugs affect the proportions of HDC and LDC differently. While "give one of the drugs" and "reduce total cholesterol" are perfectly intelligible interventions, with respect to heart attacks they are ambiguous manipulations; that is, they have differing effects in various instances, and the differences are not due to differences in other background causes, but to the fact that intervening on total cholesterol is necessarily inter-

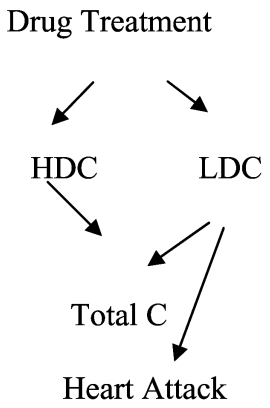

Figure 1.

vening on HDC and/or LDC, which have different effects on heart attacks. If, in contrast, HDC and LDC had the same effect on heart attack rates, interventions to alter total cholesterol would not be ambiguous.

In any case in which an identity of properties is at issue, the possibility of ambiguous manipulations—different manipulations that result in the same value of property *A* but not of property *B* with which *A* is supposedly identical—defeats property identity.

The requirement that instances of identical properties have like statistical and temporal relations is based on a simple truth: If property *A* is identical with property *B* under conditions *C*, then under conditions *C* the causes of *A* must be causes of *B*, and the effects of *A* must be the effects of *B*. That implies probabilistic connections between *A* and *B* more extensive than simply that their probabilities of occurrence in any case of *C* be equal. This consideration, which differentiates identity from epiphenomena, is the very same criterion used in ordinary science—among others, in neuroscience—for assessing causes.

I propose that mental properties and (in parallel) mental processes meeting these criteria are aggregates of comparatively microscopic physical properties and processes that, individually, may have a quite different causal role than they do collectively, in aggregation. Just why and how that could be is perhaps best understood by considering an example.

4. The Planet. A common example of "intertheoretic reduction" is the explanation of the ideal gas law by kinetic theory. For the relation of the mental and the physical the example has two appropriate features: any identifications are local, confined to gases; and there is no ultimate physical state that is identified with a temperature value—an infinity of 'micro-

states' correspond to the same temperature value. The classical identification is for equilibrium processes in which temperature does not change and the temperature plays no sequential causal role influencing other quantities. Dynamical examples, in which both microscopic and macroscopic features change over time and some macroscopic variables cause others, would seem more appropriate analogues for the phenomena of thought. A physical example is wanted that is not itself neuropsychological; climate teleconnections provide one.

Temperatures and atmospheric pressures at the surface of the sea around the globe have been recorded for more than a century—in the last 30 years or so by satellite measurements of infrared spectra. Atmospheric pressure at sea level has been recorded in the same way. Measurements in various continuous regions of the oceans vary in close connection, but correlations of these measures with one another, and with other climate phenomena, also occur among regions that are widely separated. The most famous example of such a "teleconnection" was discovered early in the twentieth century by Sir Gilbert Walker, correlating El Niño changes in the current, temperature, and pressure in the southeastern Pacific with monsoons in India. When the currents reversed direction off the coast of Chile, the monsoons failed in India.

Nowadays, regional sea surface temperatures and pressures are aggregated into climate indices with resulting distant correlations or teleconnections. The atmospheric teleconnections are produced by winds—the motions of air molecules—and, more slowly, by the motions of water molecules in ocean currents and by radiative transfer. Explaining the teleconnections from fundamental physical principles requires general climate models with thousands upon thousands of variables. And yet the teleconnections of ocean indices have a very simple macroscopic structure, in which some indices screen off others. Here are some of the principal standard ocean indices:

- QBO (Quasi Biennial Oscillation): Regular variation of zonal stratospheric winds above the equator.
- SOI (Southern Oscillation): Sea-level pressure (SLP) anomalies between Darwin and Tahiti.
- WP (Western Pacific): Low-frequency temporal function of the 'zonal dipole' SLP spatial pattern over the North Pacific.
- PDO (Pacific Decadal Oscillation): Leading principal component of monthly sea surface temperature (SST) anomalies in the North Pacific Ocean, poleward of 20° N.
- AO (Arctic Oscillation): First principal component of SLP poleward of 20° N.

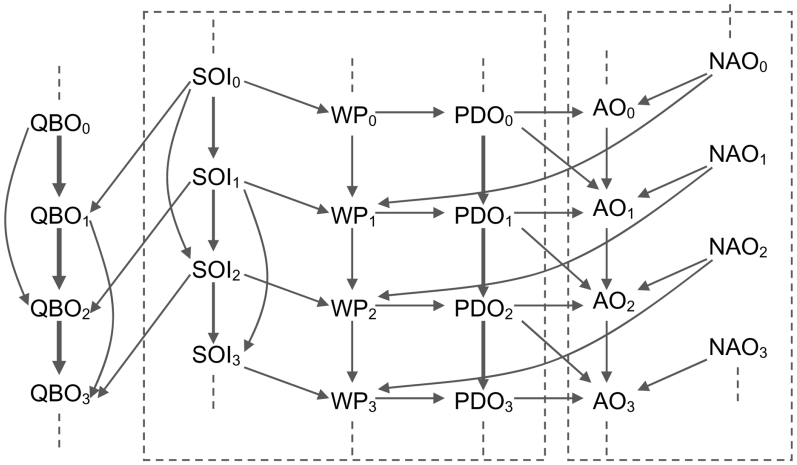

Figure 2.

- NAO (North Atlantic Oscillation) Normalized SLP differences between Ponta Delgada, Azores, and Stykkisholmur, Iceland.

The southern oscillation and other variables are not functions of features of any particular set of objects, but rather of features of whatever objects occupy a certain volume of space; and those objects and the values of their relevant variables are continually changing. The variables are recorded for hundreds of months, forming time series in which each variable is indexed by each month. Each time series for each variable can be used to generate "lagged" corresponding time series, by replacing index j with $j + n$. When the correlations of all these time series, including the lagged series, are analyzed, the result is Figure 2, from Chu and Glymour (in press).

Despite the fact that the indices do not determine the microstate of a region, the indices screen one another off exactly as in a causal sequence: the southern oscillation is independent of the Pacific decadal oscillation conditional on the spatially and temporally intermediate Western Pacific measure; WP_t is independent of SOI_{t-1} conditional on SOI_t and WP_{t-1}. These independence relations are exactly what we should expect if the arrows in the diagram represent relatively direct causal inferences and if there are no significant unobserved common causes of represented variables. Indeed, the independences are necessary if each vertex in the graph has a probability distribution that is a function of its direct sources in the graph, and there are no unrepresented sources of covariance.

General, sufficient conditions for screening off relations among vari-

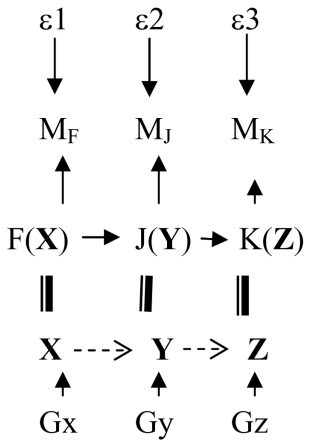

Figure 3.

ables that are identically functions of other variables can be given using a graphical criterion (Pearl 1988), but I will give only an example. Let **X**, **Y**, and **Z** be sets, or vectors, of variables, and let $F(\mathbf{X})$ be a quantity whose values are determined uniquely by the set of values of members of **X**, and analogously for $J(\mathbf{Y})$ and $K(\mathbf{Z})$. Suppose that the individual members of **X** either have no influence on members of **Y** and members of **Z**, or, if they influence **Z**, do so only through **Y**. Members of **Z** do not influence members of **Y**, and members of **Y** or of **Z** do not influence members of **X**. Consider a causal structure of the form in Figure 3, where the bars without arrows indicate collective properties uniquely determined by the collective values of members of **X**, **Y**, and **Z**, respectively. $K(\mathbf{Z})$ is a (generally indeterministic) function of $J(\mathbf{Y})$ and $J(\mathbf{Y})$ is a (generally indeterministic) function of $F(\mathbf{X})$, and the ε variables are independent sources of variation. M_F, M_J, and M_K are the measured values of F, J, and Y, respectively. The dashed arrows indicate influences by individual components of **X**, for example, that are individually insignificant. Their collective effects are the solid arrows from $F(\mathbf{X})$ to $J(\mathbf{Y})$ to $K(\mathbf{Z})$. To manipulate $F(\mathbf{X})$, for example, is to manipulate **X** at the same time, generally in any of many possible ways; to manipulate **X** is to manipulate $F(\mathbf{X})$ in some unique way. It follows that, if the variances in the ε variables are small, M_F is approximately independent of M_K conditional on M_J. We have screening off. That is what we seem to find in the climate example.

It seems plausible that what is going on with the planet's climate indices is as follows: The indices are unknown deterministic functions of underlying variables and the aggregated variable (e.g., temperature) is a function

of microvariables of the kind described above. There are ε variables, representing measurement error principally, but their variance is comparatively small. Individually, the underlying variables (e.g., the particle energies in a region) in one region have trivial influences, or none at all, on the underlying variables (the particle energies in another region), but significant aggregate influences. The aggregate influences are causal, quite as much as the individual factors they aggregate, but with a different role: A team of men may pull a wagon that no individual man can pull. Each man is a causal factor in the movement of the wagon, but a replaceable causal factor, and it is the aggregate of effort that moves the wagon. So it is with climate indices and molecular energies.

The climate network is a description of "causal roles" of the various variable types and their particular instances. The causal role of a system of macroscopic properties is the conditional independence graph, or diagram, of an aggregation of microscopic properties, together with the values of any causally relevant parameters; each macroscopic property is an unknown function of the collection of microscopic properties. The relations among an index at one time and another index at another time are stochastic, not deterministic. The value of an index is subject to external manipulation—by the sun, by human intervention, whatever—but only through the aggregate effect of the manipulation of the energy of particles and radiation in a space-time region.[8]

5. Graphing the Brain. Brain events are now measured by a variety of imaging techniques, of which nuclear magnetic resonance imaging is perhaps the best known and most popular. With the technique, the contents of some kinds of thought processes can be matched to a distinctive image

8. It is important for the analogy developed later to know that not every way of aggregating to form macroscopic variables will yield screening off relations that reflect a causal structure; indeed, most ways will not. To continue the climate example, we might, e.g., have measured the aggregate temperature and pressure differently. Computer scientists who work on data mining have developed a variety of algorithms for forming new variables by clustering cases or aggregating variables. Some of those techniques have been used to develop climate indices alternative to the conventional indices of Figure 1, developed by climate researchers over many years. A recent paper (Steinbach et al. 2003), e.g., proposes more than 100 regional ocean temperature indices and nearly 800 sea-level pressure indices. The time series of these indices do not, however, generally form robust structures like those of Figure 2 for which a connection that occurs between an index at time k and another index at time $k + 1$ occurs again at times $k + 1$ and $k + 2$, respectively, and so on. That stability does not happen with most of the computer scientists' climate indices: there is no stable screening off; the conditional independence relations are unstable and change over relatively brief times. The automated clusters are functions of the underlying energies all right, but the wrong functions in the wrong places.

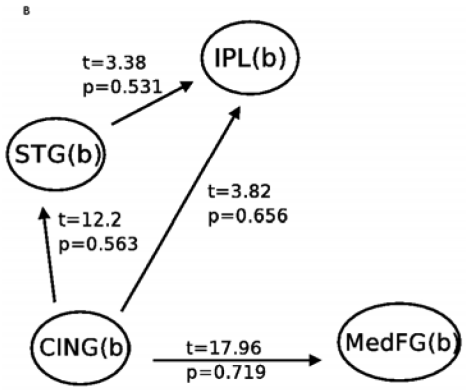

Figure 4.

pattern on regions of the brain (Suppes et al. 1997, 1998, 1999; Suppes and Han 2000; Mitchell et al. 2003). As a further step, screening off relations and graphical causal models can be developed relating kinds of events in different brain regions so that the entire neural process is associated with a kind of mental process. That has recently been attempted by several research groups independently (Hanson et al. 2006; Haxby et al. 2006; Keibel et al. 2006) using magnetic resonance images of very small brain regions to argue that one or another psychological state or process is produced by—or just *is* in the relevant individuals—a causal process among these regions. Hanson et al., for example, use magnetic resonance results on a number of brains to produce Figure 4, which diagrams influences among five brain regions in a complex experiment requiring subjects to identify "significant event changes" in a stimulus series. (IPL is the inferior parietal globule, STG is the superior temporal gyrus, MedFG is the middle frontal gyrus, and CING is the cingulated gyrus.) The structure implies (indeed, is in part obtained from) a pattern of statistical constraints exhibited by the measurements, in particular that MedFG is independent of the other variables conditional on CING.

When given a time-series representation, the results of functional magnetic resonance may look structurally very much like the graphs of time series for climate indices. For example, the graphical model in Figure 5, which is from unpublished data (Hanson et al. 2007), measures regions of the middle occipital gyrus (mog), inferior parietal lobule (ipl), middle frontal gyrus (mfg), and inferior frontal gyrus (ifg) from "six brains" watching the same video. It bears comparison with Figure 2.

A challenging next step in neuropsychology is robustly to correlate sequences of steps in cognitive tasks with sequences of regional brain

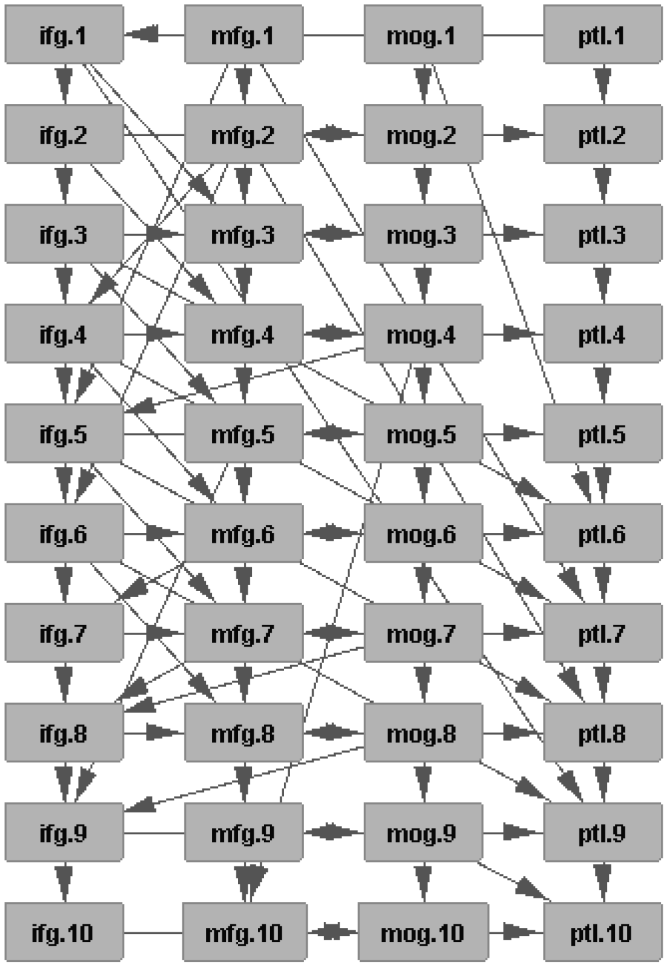

Figure 5.

activity, or sequences of modular causal processes like those in Figure 4. As far as I know, that has not been done, but it has been proposed (Poldrack et al. 2006).

Superficially, the causal hypotheses now emerging from imaging studies may seem very different from the proposals of Hawkins and Kandel, but they are structurally similar. The knowledge of physical detail is of course much more limited in the imaging studies, but that is beside the point I am pressing. In both cases, physical mechanisms are proposed for simple

cognitive processes and are conjectured to be components of more complex processes. In both cases, a detailed correlation is required; in both cases, unambiguous manipulations are sought in order to secure correct identifications. In imaging studies the existence of unambiguous manipulation is a—one might say *the*—critical matter since irrelevant brain regions may be active and wrongly selected as "regions of interest" in empirical studies. But that a goal ascribed to practitioners is uncertain of achievement does not argue that their intent is misunderstood.

6. When Is a Brain Like the Planet? So let it be with mental causation: Microscopic physical processes combine to produce a cornucopia of possible thoughts; each mental process is an aggregate of physical processes. The causal role of a kind of thought or thought process is the causal sequences of the aggregated physics it is, and the identifications involved are local, not identities in every conceivable possible world or circumstance. A philosopher can refuse to acknowledge such identifications as they are found, but she may as well refuse to acknowledge that light is electromagnetic radiation and that the mean kinetic energy of an equilibrium gas is its temperature, for those identifications are made on analogous grounds. Issues of qualia (not lightly) aside, thoughts that are of a kind that form sequences followed by other thoughts or actions of a kind, with the probability and screening off relations of aggregated physical states, are properly regarded as causal because they literally *are* the aggregated physical states and the thought processes literally are physical processes of the aggregates. Just as with the earth's climate, the identifications are local, the physical sequences need not be invariable or deterministic, and the local relations among features constitute a network of up and down identities and sideways causes.

When it thinks, a brain is like the planet.

REFERENCES

Bergson, H. ([1911] 1998), *Creative Evolution*, translated by Arthur Mitchell. New York: Dover.

Bickle, J. (1998), *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Block, N., and Robert Stalnaker (1999), "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review* 108: 1–46.

Chalmers, D., ed. (2002), *Philosophy of Mind.* Oxford: Oxford University Press.

Chu, T., and C. Glymour (in press), "Semi-parametric Causal Inference for Non-linear Time Series Data", technical report, Laboratory for Symbolic Computation, Carnegie Mellon University.

Davidson, D. (1970), "Mental Events", in L. Foster and J. Swanson (eds.), *Experience and Theory*. New York: Humanities Press, 79–101.

Hanson, S. J., et al. (2006), "Bottom-Up and Top-Down Brain Pathways Underlying Comprehension of Everyday Visual Action", working paper, Rutgers University Conference on fMRI.

——— (2007), "Bottom-Up and Top-Down Brain Functional Connectivity Underlying

Comprehension of Everyday Visual Action", *Journal of Brain Function and Structure*, submitted.
Hawkins, R., and E. Kandel (1984), "Is There a Cell-Biological Alphabet for Simple Forms of Learning?", *Psychological Review* 91: 375–391.
Haxby, J., et al. (2006), "Distributed Patterns of Activity in fMRI Data: What Do They Reveal about Neural Representations?", working paper, Rutgers University Conference on fMRI.
Heil, J., ed. (2004), *Philosophy of Mind*. Oxford: Oxford University Press.
Hill, C. (1991), *Sensations*. Cambridge: Cambridge University Press.
Kandel, E., and R. Hawkins (1992), "The Biological Basis of Learning and Individuality", *Scientific American* 267: 78–86.
Kiebel, S., et al. (2006), "Dynamic Causal Modeling for fMRI: Extending the Generative Model", working paper, Rutgers University Conference on fMRI.
Kim, J. (1993), *Supervenience and Mind*. Cambridge: Cambridge University Press.
——— (1998), *Mind in a Physical World*. Cambridge, MA: MIT Press.
——— (2005), *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
Kripke, S. A. (1980), *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Libet, B. (2004), *Mind Time: The Temporal Factor in Consciousness (Perspectives in Cognitive Neuroscience)*. Cambridge, MA: Harvard University Press.
McLaughlin, B. (2001), "In Defense of New Wave Materialism: A Response to Horgan and Tienson", in C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 319–330.
Mitchell, T., et al. (2003), "Classifying Instantaneous Cognitive States from fMRI Data", *American Medical Informatics Association Annual Symposium,* 465–469.
Nagel, T. (1974), "What Is It Like to Be a Bat?", *Philosophical Review* 83: 435–450.
Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.
Poldrack, R., et al. (2006), "How Can Neuroimaging Relate Cognitive and Neural Processes?", working paper, Rutgers University Conference on fMRI.
Scheines, R., and Spirtes, P. (2002), "Causal Inference of Ambiguous Manipulations", *Proceedings of the 18th Annual Meeting of the Philosophy of Science Association*, http://philsci-archive.pitt.edu/view/confandvol/200204.html.
Smuts, J. (1973), *Holism and Evolution*. Westport, CT: Greenwood.
Steinbach, M., et al. (2003), "Discovery of Climate Indices Using Clustering", *Proceedings of the 2003 Conference on Knowledge Discovery and Data Mining*.
Suppes, P., et al. (1997), "Brain-Wave Recognition of Words", *Proceedings of the National Academy of Sciences* 94: 14965–14969.
——— (1998), "Brain-Wave Recognition of Sentences", *Proceedings of the National Academy of Sciences* 95: 15861–15866.
——— (1999), "Invariance between Subjects of Brain Wave Representations of Language", *Proceedings of the National Academy of Sciences* 96: 12953–12958.
Suppes, P., and B. Han (2000), "Brain-Wave Representation of Words by Superposition of a Few Sine Waves", *Proceedings of the National Academy of Sciences* 97: 8738–8743.
Wegner, D. (2003), *The Illusion of Conscious Will*. Cambridge, MA: Harvard University Press.
Woodward, J. (2003), *Making Things Happen*. New York: Oxford University Press.