In *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl* (pp. 103-180), Paul K. Feyerabend and Grover Maxwell, editors. Minneapolis, MN: University of Minnesota Press, 1966.

The Compleat Autocerebroscopist: A Thought-Experiment on Professor Feigl's Mind-Body Identity Thesis

PAUL E. MEEHL

Professor Feigl's mind-body identity thesis, which may be characterized as a daring hypothesis of "empirical metaphysics," asserts that human raw-feel events are literally and numerically identical with certain physical brain-events. By physical$_2$ events he means, adopting the terminology of Meehl and Sellars (1956), events which can be exhaustively described in a language sufficient to describe everything that exists and happens in a universe devoid of organic life. Given the set of descriptive terms (predicates and functors) which would be capable of describing without residue all continuants and occurrents in an inorganic world (say, perhaps, our world in the pre-Cambrian period), we need not supplement this conceptual equipment in order to describe everything that exists and happens when a human being experiences a red sensation ("has a red raw feel," is the locus of a red-qualitied phenomenal event).

It is not my purpose here to attack or defend this identity thesis, concerning which I have, in fact, no settled opinion. Rather I hope to clarify its meaning by examining some implications and alternatives presented when we apply it to a plausible thought-experiment involving Professor Feigl's "autocerebroscope" (1958, p. 456).

I. The Thought-Experiment

The autocerebroscope is an imaginary device, differing only in arrangement and technological development from instruments already used by psychologists and neurophysiologists, which would enable a subject to receive continuous and nearly instantaneous information regarding the momentary physical$_2$ state of his own brain while he is experiencing raw feels. For simplicity of exposition as well as avoidance of irrelevant substantive problems (e.g., vagueness in the applicability of certain phenomenal predicates), I shall consider the case of two clearly distinguished color qualities, *red* and *green*. The subject is a nonaphasic, nonpsychotic, cooperative, English-speaking neurophysiologist thoroughly pretested for possession of normal color vision. To avoid the terrible difficulties of methodological behaviorism and to highlight his epistemological dilemma, it will be convenient to speak of this subject in the first person; and I invite the reader to read what follows taking himself as the "I" so spoken of.

The apparatus consists of instruments leading directly off my cerebral cortex which convert the pattern of my current brain activity into either of two symbolic patterns, R (for "red") and G (for "green"), presented on a television screen directly in front of me. The apparatus is wired and adjusted so that symbol R appears whenever the physical$_2$ state of my

visual-perceptual cortical area is that known (by Utopian neurophysiology) (a) to be produced by retinal inputs of red light waves in persons with normal color vision, and (b) to ordinarily produce a tokening of "red" in cooperative English-speaking subjects. This brain-state we designate by the lower-case italic letter *r*; also by "Φ_r."

The same conditions hold, *mutatis mutandis,* for the television symbol G in relation to the green brain-state (= *g*). It is important to be clear that the apparatus's symbol presentation depends solely upon the *visual*-cortical state, and is not wired so as to be directly influenced by other events. If, while the visual cortex is in state *r*, the cerebral tokening mechanism should happen (for whatever reason) to token "green," or the light waves entering the eye should be in the physical$_2$-green spectral region, these events are without causal influence upon the television symbol presentation.

However, the television system of the apparatus is also (and independently) arranged so as to vary the hue of whatever symbol is being presented, and the symbol (regardless of whether it is R or G in type) is sometimes colored red and sometimes green, this coloration persisting from, say, 5 to 20 seconds before it changes, and the variable interval lengths being randomly determined.

Under these conditions it is psychologically possible for me as subject to attend simultaneously to two rather simple aspects of my momentary visual experience, i.e., the *shape* and the *hue* of the presented symbol. By this arrangement we avoid the old stomach-ache about when introspection is "really" short-term retrospection, and at least minimize the touchy problem of how many things a subject can attend to simultaneously.

What instructions are given to me as subject in this Utopian experiment? We first agree that I am to use the words "red" and "green" to designate *experienced color qualities*. For simplicity and speed of reporting, the single hue-quality word "red" (or "green") will be conventionally taken as elliptical for "I am now experiencing a visual raw feel of red (or green) hue quality," respectively. As a neurophysiologist I am, of course, aware of the fact that I, and other normal English-speakers, have historically acquired our shared color language by a rather complicated process of social learning. Since the thing-predicate "red" is employed ambiguously in vulgar speech, referring indiscriminately to (a) surface physico-chemical properties of external objects, (b) the distribution of light waves such objects are disposed to reflect in "standard" illumination conditions, and (c) the raw-feel quality normally produced in persons exposed to such stimulation, a terminological stipulation is required for more exacting experimental purposes. In common life, the ambiguity is not often a source of malcommunication because either the three conditions are simultaneously realized or the context makes the speaker's intention clear. In special cases, as when a person asserts "That's red" in a red-illuminated darkroom, it may be necessary to question him as to his intention, and decide for the truth or falsity of his claim accordingly. In the present context, I as subject am instructed to employ the color predicate "red" in its phenomenal usage, i.e., as a predicate descriptive of the experienced raw-feel hue-quality.

If all goes as expected, I find myself tokening "red" as descriptive of my experienced color whenever the television symbol has the shape R, and tokening "green" when it presents the symbol G. In speaking the phenomenal language, it seems clearly appropriate to token "red" (labeling the experienced hue) and "R" (labeling the experienced sign-shape) quasi-simultaneously. If necessary, I can be instructed to report the experienced hue orally and the experienced sign-shape by depressing one of two keys. In speaking the physical thing-language, with the intention of denoting the external physical$_2$ properties of the distal

stimulus, I find myself willing (on the *basis* of my raw feel but with *reference* to the objective television screen) to allege that the screen is displaying a red-colored letter R, or a green-colored letter G, as the case may be. The received laws of psychophysiology seem to be instantiated, and I am so far content with the status of my mental health as well as the soundness of Utopian neurophysiology. Both my brain and the autocerebroscope appear to be functioning satisfactorily.

But now a frightening aberration unexpectedly occurs. One day, after an otherwise "normal" run has gone on for several minutes, I find that I am experiencing green when the (apparently green-hued) symbol is R-shaped. I announce this disparity to my research assistant, who takes a long look himself and embarrassedly informs me that the R symbol looks appropriately red to *him.* Perhaps, he suggests timidly, I am mis-speaking myself? I take stock carefully, and remain very sure that I am at the moment experiencing a green-hued phenomenal event. I find no difficulty articulating the *word* "red," and by shutting my eyes I find that I am (being an excellent color visualizer) able voluntarily to call up a pretty good red-hued image. But upon opening my eyes, I continue to "feel certain" that the raw-feel predicate "green" is the correct description of what I see when looking at the screen. I am not in the state of a severe compulsive patient who feels involuntarily impelled to utter the "wrong word," nor of the aphasic who hears himself speak a word he knows isn't what he helplessly wants to say. The word "green" seems perfectly appropriate to me, given the instructions to describe my momentary raw-feel experience. The concurrent "inappropriateness" feeling arises not from doubt or vagueness about the phenomenal quality, but solely from my scientific knowledge about the causal network in which I and my experiences are presumably embedded. The autocerebroscope informs me that my visual cortex is in the *r*-state, which it *should* be because the television screen is (according to my assistant) emitting red light. I can, of course, dispense with the assistant's report and substitute other multiple observers, plus additional physical apparatus, to confirm that my objective retinal input is physically red light.

Suppose now that I carry out detailed and thorough checking and testing of the entire apparatus, as well as gathering the testimony of multiple observers. And suppose that all of the obtainable evidence continues to indicate, via numerous tightly knit nomological relationships and over-determined particulars, that there do occur unpredictable but repeated occasions in which my visual cortex is in physical$_2$ state *r*, that state being causally determined by objective inputs of physical$_2$-red light from the physical$_2$-red-colored screen, and yet I am, on those occasions, momentarily certain that I am experiencing phenomenal green. Over an extended series of trials, this happens in about 10 per cent of my phenomenal reports. A critical question here is, of course, what precisely this "momentary certainty" amounts to (e.g., is it more than the fact of psychological indubitability?). We shall set this question aside for now, because we can approach it more fruitfully after examining some further ramifications of the autocerebroscopic thought-experiment in its causal-scientific aspect. Proceeding still at a "common-sense" level (Utopian-science common sense, that is), how do I meta-talk about my situation as aberrant autocerebroscopic subject? I want to token "I experience green" in obedience to the instructions, and I do not need to infer this raw-feel proposition from any other propositions. If my colleagues ask me *why* I persist in saying "I experience green," I can only remind them of their instructions, which require me to describe my phenomenal events, rather than to make causal claims about the apparatus or the physical$_2$ state of my visual cortex. If asked whether I consider "I experience green"

to be incorrigible, I reply that all empirical propositions are corrigible—perhaps adding that if I keep my sanity I may end up being forced to correct "I experience green" in particular! However, I confess that I can't cook up a genuine *doubt* that I am now experiencing green. This means that while I sincerely admit corrigibility, I am at the moment unable "seriously to entertain the notion" that the future evidence will, in fact, turn out that way. I know in general what such evidence would consist of, and I firmly believe that it *would* be rational of me, in the face of such evidence, to conclude that my presently tokened "I experience green" was false. I hope that I am a rational enough man so that I *would* at such time find it psychologically possible to abandon the proposition. And if I were not to turn out in the event to be that rational, still I am prepared *now* to say that such a development would prove me to be less rational than I had supposed. Nevertheless, I am not now psychologically able to make myself seriously doubt that I am experiencing green, and hence I don't seriously entertain the notion that the situation will arise. In short, I remain confident that, if we keep looking, we will succeed in locating the "bug" in the autocerebroscope which, I hypothesize, is the true explanation of these crazy events.

Is there any dishonesty, unreasonableness, or inconsistency in this combination of assertions and expectations? If I am a Utopian identity theorist, I consider it physical$_2$-impossible to "experience a green raw feel" while my brain is in state *r*. To assert such a thing would be like saying "This soup, while very hot, consists of motionless molecules." Such a statement is frame-analytically[1] false within the nomological network of physics; one who holds the kinetic theory of heat might go so far as to call such a statement meaningless.

Suppose that exhaustive tests of all conceivable kinds fail to reveal any defect in the autocerebroscope's structure, function, or brain-attachments. At some point I become convinced that there are times when my visual cortex is in state *r* but I am simultaneously tokening "I see green," and that this (inner) tokening "seems clearly descriptive of my raw feel." What are the possibilities open to me for making *causal* sense of such a bizarre state of affairs?

We begin with the received doctrine of Utopian neurophysiology, which accepts the identity thesis and which further identifies a particular brain region (or, better, system of related cell-assemblies) as the physical$_2$ locus of events whose occurrence *constitutes* a visual raw-feel event. (I believe that Professor Feigl is clearly committed, although he is not very happy about it, to saying that a raw-feel event is literally, in a physical$_2$ sense, *in the head*—since otherwise he contradicts Leibniz's Principle. See Section IV below.)

---

[1] Throughout this paper I use the expression "frame-analytic" to mean, roughly, true by "theoretical definition"; which latter phrase in turn means, roughly, stipulation of meaning (explicit or implicit) in terms of other theoretical constructs which are themselves "defined implicitly" by the accepted nomological network. While such frame-analytic truths therefore rest in one sense upon conventions, these conventions are far from "arbitrary," but are adopted on the basis of our theoretical knowledge—our current best available notion of "how the world is." The deeper issues raised here (e.g., status of so-called conventions in empirical science, clarity and defensibility of the traditional analytic-synthetic distinction) are beyond the scope of this paper and of my competence. Frame-analyticity is closely related to truth by P-rules, by meaning-postulates or A-rules, and the like. See, for example, Carnap (1950, 1952), Maxwell (1961), Sellars (1948, 1953), Feyerabend (1962). I do presuppose in employing the phrase "frame-analytic" that whatever may be the final resolution of this cluster of technical philosophical problems, *some* important distinction will be preserved between the kinds of analyticity involved in "bachelor = unmarried male" and "temperature = mean kinetic energy of molecules."

Presumably Utopian psychophysiology asserts—or, better, for one who accepts its nomological network, "implicitly defines"—a one-many relationship between raw-feel predicates and physical$_2$ brain-state functors. A set of structural assertions about neurons (numbers, positions, and synaptic connections of a very complex kind) identifies the cerebral system which is the *locus* of visual raw-feel events; thus, visual raw-feel events cannot occur in the transverse gyrus of Heschl (auditory projection area), but they can occur in the calcarine cortex (or Brodmann's area 18?). We designate by "V" the cortical region or functional subsystem which is the locus of visual raw-feel events. Given an appropriately structured cerebral subsystem, its momentary state is exhaustively characterized by a set of physical$_2$ functors. These might be simple (e.g., strength of electromagnetic and electrostatic fields), or, more likely, complex (e.g., second time-derivative of a proportion of instantaneously activated synaptic knobs on cells of type X in cell-assemblies of structure S). The received neurophysiological network asserts that a necessary and sufficient condition for experiencing a red raw feel (or, the theoretical *definition*, within this causal framework, of brain-state *r*) is that a cerebral system of type S must be in a state described by a complex conjunction of physical$_2$ functor inequalities:

P: One-place predicate designating the phenomenal quality,
L: Two-place predicate locating a brain-event in the brain of a person,
Ψ: Two-place predicate designating the internal *Erlebnis*-relation, "...experiences phenomenal event...,"
Φ: One-place predicate designating the complex physical$_2$ property which a brain-event has when its physical$_2$ functors satisfy certain inequalities (the relation of raw feel to brain-state being, presumably, one-many),
x: Variable ranging over persons,
y: Variable ranging over phenomenal events,
z: Variable ranging over physical$_2$ brain-events.

Reference to time is omitted, taken as quasi-simultaneous.
Then the empirical "psychophysical correlation-laws" are:

$$(x,y)\,\Psi\,(x,y) \bullet P\,(y) \rightarrow (E!z)\,L\,(z,x) \bullet \Phi\,(z)$$
$$(x,z)\,L\,(z,x) \bullet \Phi\,(z) \rightarrow (E!y)\,\Psi\,(x,y) \bullet P\,(y).$$

The Feigl theory consists of conjoining to each of these correlation-laws the identity-assertion

$$(y = z).$$

But we have so far not done justice to the advanced development of Utopian neurophysiology. Although our thought-experiment began by wiring and attaching the autocerebroscope only to provide information about events occurring in the cerebral locus of visual raw-feel events, Utopian knowledge of brain function includes much besides this. For the "normally functioning brain," we also possess scientific understanding of the causal relations obtaining in other cerebral systems, including the tokening mechanism. This means that certain problems of methodological behaviorism, and certain philosophical difficulties arising from reliance upon vulgar speech, have been "solved"—insofar as empirical knowledge ever solves problems. The differences between raw-feel utterances which are "correct," "false because of lying," "false because of mis-speaking," "false because of being hypnotized," "false because of aphasia," "false because of slovenly language training," "false

because of having previously misread an English-German dictionary," etc., are formulable by reference to *where* in the intracerebral causal chain the tokening process and its controls have gone awry. Presumably Utopian neurophysiology will have isolated a cerebral system T (= the tokening system) which is the physical locus of events t_r, t_g, etc., these events being the inner tokenings of raw-feel predicates "red," "green," etc. These tokening events are the immediate causal descendants of raw-feel events in the visual system V; and they are the immediate causal ancestors of events in intermediate instrumental systems which arouse, trigger, and monitor subsequent motor-control systems that give rise to families of overt acts of the reporting kind (vocalizing "red" or pressing a red-colored lever). Detailed experimental and clinical analysis will have made clear which system does which, and it will be precisely known how, for example, a red-qualitied visual raw feel gives rise to vocalizations of "red," "rot," or "rouge" in a trilingual subject, depending upon the instructions given him or his perception of the momentary social context. Obviously none of the three cell-assembly systems on the motoric side which control *utterances* of "red," "rot," or "rouge" would be considered as the primary tokening mechanism, especially for purposes of philosophical discussion. We need not here decide upon the precise conditions necessary for identifying the primary tokening system T, since for our purposes it will suffice to place certain conditions upon it. It must at least undergo states which are physical$_2$-distinguishable, and these distinguishable states must be correlated (in a normal person) on the one side with the raw-feel events, and on the other with "appropriate" states in the first-order motoric system. That is, T is the locus of events t_r and t_g which are the causal descendants of raw-feel states *r* and *g* respectively; and the states t_r and t_g are the causal ancestors of events m_r and m_g respectively, these latter being events in the "English-set motor-control system" which—*ceteris paribus*—give rise to differentiated chains continuing through the motor area to effector-organ events (vocalizations of "red" or "green"). In this scheme, the tokening mechanism T is the physical$_2$ locus of tokening *propositions* (= "making judgments"), whereas the events *m* are tokenings of sentences. For present purposes, a subject tokens "red" when the primary tokening system T is the locus of physical$_2$-event t_r regardless of whether he utters, or "tends to utter" (by covert laryngeal twitches) the English word "red" or the French word "rouge," or even if the process is for some reason stopped short of affecting any part of the instrumental reporting mechanism.[2] What we require, in short, is that system T must be the locus of physical$_2$-differentiable events t_r and t_g with input and output conditions appropriately correlated. It will not suffice, for example, to find a system which is activated whenever a visual raw-feel event *recurs*, and whose correlated report is one of mere "familiarity" (e.g., "I have experienced this color before"). When the hue of a raw feel is our subject matter, the primary tokening mechanism must be the locus of distinguishable symbolic events that are hue-correlated.

The nomologicals of Utopian neurophysiology not only assert causal dependencies between raw-feel events in V and tokenings in T (e.g., $r \twoheadrightarrow t_r$, $g \twoheadrightarrow t_g$) but they also

---

[2] This formulation does not, I would think, prejudice the philosophical issues, and is simpler to talk about for present purposes. If no such mediating judgmental tokening occurs, then for the "propositional," primary tokening event *t* we would presumably have to substitute some sort of conjunction of (1) an "English-set" superordinate event, elicited by one's perception of the audience as being English-speaking, and (2) the first link in an English-verbalizing event-sequence, which link is activated (instead of French or German) because of the superordinate "English-set" regnancy. These are presently unsettled issues in psycholinguistics.

permit these nomologicals to be derived as theorems. That is, the structural statements about how the brain is organized genetically, when combined with more fundamental laws of neurochemistry and physics, suffice to explain neurophysiological laws of such intermediate molarity as ($g \twoheadrightarrow t_g$). Gross (and merely stochastic) regularities at the level of molar behavior (e.g., "Normal people almost always report 'red' as an afterimage of green stimulus inputs") are shown to be physical$_2$-deducible from a combination of neurophysiological laws of intermediate level with detailed narration of social learning histories. These intermediate-level laws are themselves deducible from structural laws about how the human brain is wired, together with microlaws expressed in terms of microanatomy, biochemistry, and physics. The stochastic character of the more molar laws is itself explained within the system, and provides a causal account of the vagueness intrinsic to most ordinary phenomenal predicates.

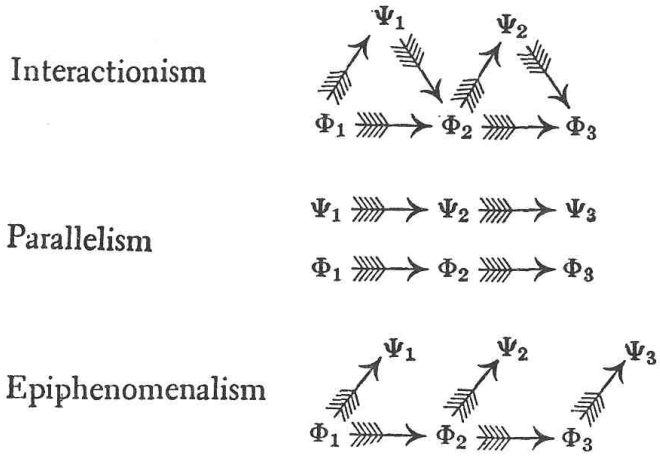
If Utopian neurophysiology embodies Feigl's identity thesis, it does so on the basis of much more evidence than that available to Feigl in the mid-1960s. Why has Utopian neurophysiology augmented its psychophysical correlation-law with the conjoined identity statement? Why has it not preferred a psychophysical interactionist nomological of form $g \twoheadrightarrow \Psi_g \twoheadrightarrow t_g$ which speaks of a raw-feel event Ψ_g that is not physical$_2$ but only physical$_1$ (belonging in the space-time network)?

I submit that the reason for this scientific decision would not be different from any other option in which the scientist dispenses with a supernumerary event or entity in concocting and corroborating his causal picture of the world. He does not feel under any obligation to "rule out" alleged events Ψ_g, but only to show that they are causally dispensable. His situation is like that of the geneticist, who began with Mendelian "factors" (evidenced on the molar side by certain phenotypic breeding statistics) and ultimately *identifies* them with genes (i.e., with chemical packets found in specific chromosomal loci). We do not ask the geneticist to "prove that there aren't factors 'associated with' genes," once he has shown that the causal role played by factors in explaining the phenotypic statistics is indistinguishable from the causal role assigned to genes in physiological genetics.

I am not here attempting to beg the crucial philosophical question, whether it can even *make sense* to identify a simple phenomenal predicate's intension with the meaning of a complex physical expression. Our Utopian neurophysiologist may be guilty of a philosophical mistake, through not seeing that the mind-body problem involves semantic (or epistemological?) issues which are unique. This question we shall examine below. My point here is that *if* it is philosophically admissible to assert the identity thesis, *if* it can be considered meaningful at all, then the empirical grounds for embodying it in the network will be of the usual scientific kind. The scientific aim is to concoct a nomological network in which all events find their place; if the causal antecedents and consequences of Ψ_g are indistinguishable from those of g, we have merely two notations for one and the same scientific concept. The abstract possibility that Ψ_g should be retained to designate a kind of event distinct from g which *might conceivably exist* "alongside it," but lacking any causal efficacy, would not impress any scientist. He would properly point out that when *Lactobacillus bulgaricus* was shown to be the causative agent behind the curdling of milk, the lowliest *Hausfrau* ceased to speak of brownies; that geneticists do not hypothesize factors "parallel with" or "mediating the action of" genes; and that caloric has been dropped from physical vocabulary since the kinetic theory of heat. There "could still be" such things as brownies, factors, and caloric; received science may be in error, and one cannot refute an existential

proposition. But when the previously assigned causal role of a hypothesized entity is found to be either unnecessary or identical with a more fully known entity, the separate existence of the old one is no longer seriously maintained.

The "possibility" of a parallel event lacking causal efficacy has a special interest in the raw-feel case, which it may be profitable to examine here. It is sometimes argued that psychophysical interactionism, psychophysical parallelism, and epiphenomenalism are meaningless pseudotheories because they could not in principle be distinguished. In the diagram we represent causal connection by the arrow $\twoheadrightarrow$, distinguishable mental events by Ψ_i, distinguishable physical$_2$ events by Φ_i, and time relationships by the correlated subscripts *i*. The issue of simultaneity versus precedence in causation is set aside for present purposes.



Critics of these distinctions point out that while the diagrams make it appear that we deal with three theories, they achieve this (misleading) effect by arbitrarily dropping selected arrows which represent perfectly good nomologicals. If Φ_1 is a necessary and sufficient condition for Φ_2 and also for Ψ_1, then the conjunction $\Phi_1 \bullet \Psi_1$ is necessary and sufficient for Φ_2; and which nomologicals we elect to draw in as "causal arrows" is surely arbitrary. Hence the three theories have indistinguishable content.

In spite of a fairly widespread acceptance of this analysis, I believe it to be mistaken, or at least in need of further justification by invoking an ancillary principle which I see no good reason to adopt. The usual practice in science, when inquiring as to the existence and direction of causal arrows, is to carry out the analysis at all levels. When an experimental separation of confluent factors is technologically (or even physically) precluded, we do not abandon our attempt to unscramble the skein of causality by distinguishing merely correlated from causal relations. The commonest method of doing this is microanalysis, which in advanced sciences is extremely powerful, and often suffices to satisfy any reasonable theoretical interest. In the biological and social sciences, where much of our evidence (especially in "field" and "clinical" studies) consists of correlations within the material as we find it, the only reasonable basis for choice may lie in moving our causal analysis to a lower level of explanation.

In the psychophysical problem, it should not be assumed that the physical$_2$ events designated by the Φ's are incapable of further analysis. If we characterize Φ_1 and Φ_2 by conjunctions of physical$_2$ expressions designating their respective microdetails, so that Φ_1 is written out as a complex conjunction of physical$_2$ functors (limiting case: elementary physi-

cal particles, electromagnetic fields), we ask whether the intermediate-level law ($\Phi_1 \twoheadrightarrow \Phi_2$) is derivable from the fundamental nomologicals (the structure-dependent features of Φ_1 being now packed into the microconjunction). If ($\Phi_1 \twoheadrightarrow \Phi_2$) is microderivable, then our causal account of Φ_2 is complete without the "mental correlate" Ψ_1 being required. The nomological relation between Ψ_1 and Φ_2 is a universal correlation but not a relation of causal dependence. If we retain Ψ_1 in our network, it will have to be for an extrascientific reason, such as a philosophical argument refuting the identity thesis analytically.

Having thus distinguished interactionism from the other two, can we tell *them* apart? Suppose that the phenomenal events themselves have sufficient richness to permit characterizing each of them by, so to speak, crude "phenomenal quasi-functors" (e.g., the color solid, the smell prism). Then we can try to formulate various intraphenomenal causal laws involving these dimensions and their combinatorial laws, and we can ask whether these combinatorial laws are in turn theorems derivable from a more basic set of Φ-Ψ laws. If they are, the usual scientific practice would be to decide for epiphenomenalism, on the ground that parallelism leaves the intraphenomenal combination-laws unexplained. Another approach would, of course, be experimental separation. The Utopian neurophysiologist can induce Φ_2 directly by imposing an artificial intracerebral stimulus, interrupting the immediately preceding events, whereby Φ_2 occurs without Φ_1—or Ψ_1!—having preceded it. If the phenomenal event Ψ_2 then occurs, we conclude that its "regular" phenomenal antecedent Ψ_1 is not part of the causal ancestry. The normal causal role of Φ_1 is unchallenged because the artificial stimulation is physical$_2$-identical with that "normally" imposed by Φ_1.

The ancillary principle alluded to above, which would be needed to defend the indistinguishability of the three psychophysical theories, is that whenever an event is time-place correlated with another, it must be taken to be causative unless shown *not* to be. I see no reason for adopting such a principle in dealing with the mind-body problem, since we do not adopt it anywhere else.

As Utopian neurophysiologist, I have adopted the identity thesis because everything I know about raw-feel events enables me to plug them into the nomological network at the place I have plugged in physical$_2$ brain-states. For ordinary purposes, I continue to use the phenomenal predicates "red" and "green," just as the heating engineer talks to a householder in terms of "winter" and "B.T.U.'s" rather than in the theoretical language of meteorology or kinetic theory. The epistemological peculiarities of raw-feel propositions may be of no interest to me, but if they are, I find it easy to explain them. By "explain," I do not of course mean that all logical and epistemological *concepts* are reducible to physical$_2$ *concepts* (compare Sellars, 1953). I mean only that, given these philosophical concepts, the fact that they apply in certain unique ways to raw-feel propositions is causally understandable. Why do I have privileged access to my raw feels? Because my tokening mechanism T which tokens propositions descriptive of my raw feels is in my head, wired directly to the locus of my raw-feel events; and this is not true of *your* tokening mechanism in relation to my raw-feel events (compare Skinner, 1945; Reichenbach, 1938, pages 225-258). Why are some raw-feel properties not further analyzable, their predicates not further "definable" in raw-feel language (the "ineffable quale")? Because the physical$_2$ components of certain raw-feel events have not been separately linked to distinguishable reactions of my tokening mechanism, and some of them cannot be so linked. Why are my raw-feel predications associated (usually!) with such subjective certainty? Because this special class of tokenings has a history of thousands of reinforcements, with near-zero failures. Why does it seem that raw feels are

immediately given, not requiring inference? Because that's how I learned to token raw-feel propositions, by a direct $g \rightarrow t_g$ transition, unmediated by other tokening events linking propositions to other propositions.

Within this causal network, my problem with the aberrant raw feels is pretty clearly defined. *Something has gone wrong between my raw-feel events* (locus in visual-cortical system V) *and my primary tokening events* (locus in tokening mechanism T). We have established by repeated experimentation that the causal sequence is running off as usual up to (and including) events in V. That is, the objectively red television screen is emitting red light which produces the normal photochemical effect in the cones of my retina, which produces the normal pattern of nerve impulses through my second cranial nerve and is relayed normally in my lateral geniculate bodies and back through my optic radiations to my visual cortex. Similarly on the instrumental (output) side, my primary tokening t_g is activating the motoric systems for the utterance "green" and my laryngeal muscles are working satisfactorily. The motivational-affective systems whose activity constitutes a "feeling of appropriateness" between my primary tokening event t_g and my overt utterance are also functioning normally, and I do not feel that I am "unable to express what I experience." In short, each system is functioning normally, in obedience to the received nomologicals, except the linkage system between V and T. The "privileged access" nomological ($r \nrightarrow t_r$) seems to have broken down.

But we are not stuck with this as a rock-bottom fact. It is a complex fact, a fact with "parts" (the event has literal, physical$_2$ parts which constitute it). Since we know that the erstwhile law ($r \nrightarrow t_r$) is a theorem, derivable from the conjunction of microstructural and microfunctional propositions descriptive of a "normal human brain," and of the events r and t_r occurring within such a brain, it follows that my brain must not be a normal brain, assuming that the fundamental nomologicals (physics) are valid.

What sort of abnormality might this be? For our thought-experimental purposes, one kind will do as well as another. If none of my other functions are impaired (e.g., affectivity, verbal reasoning, auditory discrimination, rote memory, motor coordination) it is presumably not a general biochemical defect of single-neuron function, which should produce detectable aberrations in other systems as well. If my verbal learning history has been normal, the enduring microstructural residues of color-tokening activity should be the same as other people's—if the initial micro-structure was normal. The best guess is, therefore, that my "visual-associative" system's wiring diagram was initially aberrated microstructurally, so that the imposition of a standard color-learning history upon it has yielded an acquired microstructure such that the control linkage between V and T is stochastic rather than nomological.

To carry the analysis further we must raise the question whether Utopian neurophysiology is strictly deterministic. The stochastic character of nervous activity (e.g., Lorente de No's "optional transmission" at the synapse, or the spontaneous discharge of unstimulated neurons) may be attributable to intracellular events which are intrinsically deterministic but, from the standpoint of the neurophysiologist, essentially random. In addition, it is possible that genuine quantum indeterminacy operates, considering the distances and energies involved at the synaptic interface between a single terminal knob and the cell membrane of the postsynaptic neuron (see Bohr, 1934, pp. 116-119; Eccles, 1951; 1953, pp. 271-286; Eddington, 1939, pp. 179-184; Jordan, 1955, pp. 108-113; London, 1952; Meehl et al. 1958, pp. 190-191, 214-215, 328-337). Whether one or both of these factors are responsible, the

stochastic character of spontaneous discharge and synaptic transmission may be taken as an empirical fact. Approximation of intersystem stochastic control to nomologicality can be achieved by sufficient overdetermination through involvement of large numbers of elements. Thus it is known that many, perhaps most, synaptically induced discharges are produced by presynaptic activity in excess of that needed to get the cell over threshold. It must not be forgotten, however, that a very great deal of behavior is only stochastically predictable, presumably reflecting the fact that even strong linkages may allow for low but nonzero probabilities of control failure.

The normal brain is so wired that the long-term consequence of social learning is a microstructure yielding complete control between V and T. (Query whether this is literally true. Only, I suspect, if we assume that all mislabeling of pure hues is "Freudian," which is at least debatable.) This must mean that in spite of spontaneous discharge and optional transmission, the number of neurons involved in a simple hue-tokening is so large, given their arrangement, that it yields a quasi-nomological. (We must still prefix "quasi," since if $p > 0$ for failure to transmit at each synapse, $p > 0$ for system failure. But this magnitude may be such that neurophysiology pays it no more heed than physicists pay to Eddington's ice cube heating up the warm water.) So we conclude that anyone might, theoretically, token "green" when experiencing red, without having anything structurally wrong with his brain. It follows that I *might* be the victim not of a miswired brain but of the binomial theorem. This latter, however, involves such an infinitesimal probability that we determine to accept it only as a last resort.

The obvious next step in investigating my aberrated tokening is to examine the microstructure of my tokening mechanism. My single-cell biophysics (e.g., spike amplitude, speed of transmission, afterpotentials) being established as normal, we already know that the terminal impulses arriving from V are of the appropriate kind. If necessary, this can be checked directly by microleads from these termini, by study of the micro-fields and transmitter substances there produced, etc. A plausible guess is that the number and spatial distribution of synaptic knobs, either those arriving from V or those linking the neurons into cell-assemblies within T itself, are inadequate to "overdetermine" the tokening events t_r and t_g. By convergence and divergence, a subset of "trigger" neurons in the input subsystem of T, when discharged by presynaptic activity at the termini of V-fugal fibers, normally suffices to determine t_r or t_g in T as a larger system of cell-assemblies. But if this overdetermination is insufficient, the probability becomes nonnegligible that the pattern of optional transmissions and spontaneous firing of the subset of "trigger" elements will result in the "wrong" tokening. In my case this probability has reached, let us suppose, the easily detectable value $p = .10$.

Being a zealous scientist, and considering the low risks attendant upon Utopian neurosurgery, I suggest that a biopsy be performed to corroborate the hypothesis of aberrant synaptic-knob distribution. Previous research has shown what this distribution is in a random sample of normal persons; and theoretical calculations have shown that the density and placement of knobs is such that $p < 10^{-3}$ for the average individual's tokening mechanism to discharge in pattern t_g when receiving input from a modal *r*-state. It is now noted with interest that, in the researched sample of presumed "normals," individuals lying beyond the 2σ point (some 2-3 per cent of the population) would develop p values as high as 10^{-2}; suggesting that a small but nonnegligible minority of the population are tokening hue predicates erroneously 1 per cent of the time. Since 1 per cent is still quite rare, and since the

phenomenon itself is of no clinical interest and each single occurrence likely to go unnoticed or explained away (e.g., "Freudian slip," "I'm tired"), we are not surprised to find that only the autocerebroscopic experiment, performed luckily on an unusually deviant subject, has even called it to anyone's attention.

The biopsy being performed, statistical study of the sections reveals a peculiar "thinning" and "clumping" of synaptic knobs, differing from their normal distribution. Theoretical calculation shows that my T mechanism should be expected to mistoken as between a pure red and a pure green on approximately 12 per cent of occasions under the special condition of continuous, randomly alternated retinal inputs provided by the experiment. This value differs well within combined errors of instrumentation and sampling from the observed 10 per cent in experimental runs on me to date.

Scientifically speaking, everything is again satisfactory. The data are in excellent accord with theory and the particularistic hypothesis about me. If I retain my belief in the identity thesis, I will say: "Because of a structural defect in my tokening mechanism, I token 'green' around 10 per cent of the times when I am experiencing a red raw feel, and conversely. These mistokenings of course 'seem right' at the time, because what T tokens is propositions, not English sentences; and to 'seem wrong' a mistokening must occur farther along the intracerebral causal chain, as when I can't find the right *word*. In my case, the word-tokening 'green' is the normal one for a primary tokening event t_g in T, so no tendency to feel or report a disparity occurs. It is therefore literally correct to say what many philosophers have considered nonsense, namely, sometimes my raw feels *seem to be green when they are in fact red."*

It is important in contemplating this paradoxical statement to keep in mind that until some sort of tokening-of-a-universal occurs, we cannot properly raise questions of "being right," "being sure," or "knowing." It is extraordinarily tempting to forget this, especially in dealing with raw-feel judgments. Thus, in philosophical conversation, I may imagine myself to be experiencing green, and the impulse is to say exasperatedly, "But surely I couldn't be wrong about *that*!" The trouble is that this imagery of green leads me to think that *if* I were to call *that* imaged green "green," I could not conceivably be mistaken. And if my image is green, this is certainly true, being tautologous when set forth propositionally; i.e., if I am experiencing green, it cannot be a mistake to call it "green." The trick is in the imagery, whereby I subtly introduce the hypothesis that I *am correctly labeling my experience*. Nor does this tempting error, I think, have any special relation to the identity thesis. A complete metaphysical dualist, for whom phenomenal green is a state of a nonspatial psychoid causally connected to a brain, must also realize this is a mistake and be wary of it. There simply is no necessary connection between "Jones tokens 'green' at *t"* and "Jones experiences green at *t,"* regardless of one's Jonesian ontology.

II. Empirical Character of the Identity Thesis

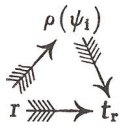
The outcome of our autocerebroscopic studies, while scientifically satisfactory, suggests a disturbing philosophical thought. Professor Feigl insists that his identity thesis, while somewhat speculative and touching on some rather metaphysical questions about the nature of things, is nevertheless a form of *empirical* metaphysics. This means that the identity thesis might, in principle, be discorroborated by scientific evidence. It occurs to us that we imagined the autocerebroscopic thought-experiment to have come out in a particular way, a way compatible with the identity thesis. Was this too easy? What sort of empirical result

could have led to our rational abandonment of it? It seems only fair, if we are dreaming up Utopian neurophysiology, to test the allegedly empirical character of the identity thesis by imagining an adverse outcome of the autocerebroscopic research. If we can't do so, something is wrong with viewing the identity thesis as an empirical claim.

To cook up an adverse empirical outcome, let us again proceed in the ordinary scientific way, by postulating a theory contradictory to the received one and deriving its consequences. Suppose there exist psychoids ("minds," "souls," "diathetes") which are substantive entities, not composed of physical$_2$ parts or substances, not space-occupying, and of such nature that most of the predicates and functors appropriate to physical$_2$ occurrents and continuants are inappropriate to them. Thus we can ask about the mass, spin, diameter, charge, etc., of physical$_2$ particles; we study the velocity, amplitude, and wave length of physical$_2$ waves; at the level of ordinary physical things, we treat of their color, shape, volume, temperature, texture, and the like. But if one were to ask about the specific gravity of an angel, we would know he had failed to grasp the idea, as when Haeckel defined God as a "gaseous vertebrate." It will be convenient, however, to adopt the convention that psychoids can be spoken of as being *at* a place in physical$_2$ space, even though they cannot *occupy* a region in the ordinary sense. This convention is perhaps dispensable, although somewhat inconveniently; but it introduces no confusion if we stipulate (as Aquinas did for angels) that "a psychoid is located where it acts" (i.e., where it exerts causal efficacy upon physical$_2$ entities). It goes without saying that psychoids must share *causality* with physical$_2$ entities (i.e., they must be physical$_1$) for us to be able to know about them. A disembodied and causally disconnected psychoid would be unknowable by us humans in the present life, as Professor Feigl has clearly shown (1958; see also Meehl, 1950). The form of ontological dualism we shall consider makes the further hypothesis that each psychoid is "connected to" an individual human brain, meaning by this that it has a bidirectional causal relation to the physical$_2$ states of one and only one brain. Finally we hypothesize that no physical$_2$ events affect the psychoid except those occurring in the brain to which it is specially connected, nor does it exert any causal influence upon other physical$_2$ events. Thus, for simplicity, we assume that clairvoyance and psychokinesis (as distinguished from telepathy) either do not exist, or are special types of brain-mediated transactions.

The psychoid is conceived to *undergo states,* which change over time and whose occurrences are the causal ancestors and descendants of specific physical$_2$ events in the brain to which each psychoid is coordinated. The existence of a psychoid and its being causally linked to a particular human brain are taken as fundamental facts of the physical$_1$ order, like the fact that there are electrons.

Suppose now that among the transitory "states" into which psychoids get are states of visual experience. When a human brain undergoes state r in its visual cortex system V, the coordinated psychoid ψ_i is causally influenced so as to experience phenomenal state ρ. This state of affairs may be designated by $\rho(\psi_i)$ and the psychophysical correlation-law can only be written properly with use of the psychoid notation. The psychoidal event $\rho(\psi_i)$ is linked by a psychophysical nomological to the physical$_2$ tokening event t_r, and we shall assume that this tokening event occurs upon the confluence of physical$_2$ inputs from V and the concurrent physical action upon T by ψ. Diagrammatically,



Utopian neurophysiology would presumably have considered this theoretical possibility even prior to discovery of the aberrant autocerebroscopic findings, and would have accepted or rejected it depending upon the microderivability of the ($r \twoheadrightarrow t_g$) quasi-nomological as a physical theorem. The theoretically expected departures from strict nomologicality would, of course, be conceptualized as due to the rare confluence of indeterminate microevents having low, but nonzero, joint probability. Prior to the aberrant autocerebroscopic findings, Utopian neurophysiology might have been erroneously (but, on the extant data, quite properly) betting on the identity thesis.

What are the logically available possibilities?

1. $r \twoheadrightarrow t_r$ holds. This, a theoretical consequence of received Utopian neurophysiology, is now refuted by the autocerebroscopic experiment. The putative (raw feel $\twoheadrightarrow$ tokening) "law" was microderivable as a physical$_2$ theorem, so the micropostulates must be modified *or supplemented by postulates concerning additional theoretical entities.*

2. $r \twoheadrightarrow_p t_r$ holds. The (raw feel $\rightarrow$ tokening) law is stochastic, and its low-probability deviations have now been brought to our attention. Under this case, two subcases are distinguishable: (a) The incidence and micropatterning of deviations are "random," and microderivable from the physical$_2$ microlaws by applying probability theory to the empirical distribution of initial microconditions. (b) The incidence and micropatterning of deviations are "systematic," and cannot be derived as in (a).

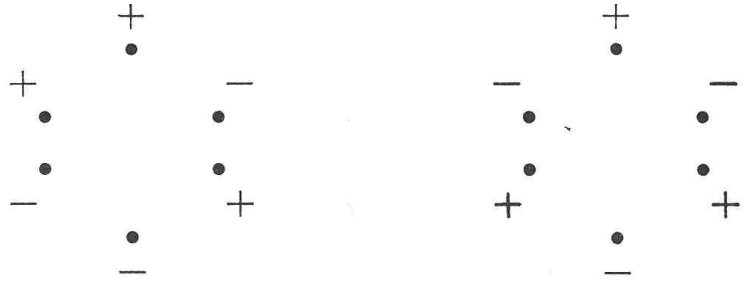
This exhausts the possibilities. Clearly (1) and (2a) are both compatible with the identity theory. The former yields a strict deterministic identity theory, the latter an indeterministic one. Neither requires postulation of additional theoretical entities mediating the empirical (raw feel $\rightarrow$ tokening) relation. The first involves strict nomologicals linking the physical$_2$ raw-feel event to the tokening event; the other involves probability linkage only, and the numerical probabilities are derivable from the microconditions.

It is in case (2a) that the psychoids can reveal themselves by exerting causal influence. Suppose it is shown that when a subset of microevents (e.g., synaptic events at individual end-feet) are locally indeterminate, i.e., the physical$_2$ microlaws make k local outcomes quantum-uncertain, the joint probability of a certain outcome pattern is p_1, another such p_2, and so on. Suppose further that the sum of all p's associated with a tokening t_g as intermediately molar outcome is $p(t_g)$. Finally, suppose that $p(t_g)$ is significantly smaller than its observed value over a sufficiently long series. Then we have discorroborated the "random" hypothesis, and may be impelled to concoct a theory embodying a "systematic selection" process determining the locally random microevents. One such theory could be the existence of psychoids, which "select" locally indeterminate outcomes *teleologically* (i.e., the psychoid "throws" a subset of indeterminate events so as to yield a patterned tokening event t_g).

It might be thought that this contradicts our hypothesis of physical$_2$ indeterminacy—that the system is either "random" or "lawful" but we can't have it both ways. This objection, while superficially plausible, is not sound. There is no contradiction between asserting that the individual microevents are "locally random," and that their *Gestalt* is systematic with respect to the molar outcome. A simple example will suffice to demonstrate this, as follows:

Consider a circular arrangement of elements, each of which "fires" on exactly half of the "occasions." Adjacent elements are wired so as to send stimulating termini to common elements in the next system, these latter elements requiring simultaneous stimulation by two inputs in order to be discharged. Hence the next system will be activated only if adjacent elements in the control system fire concurrently, but not if only nonadjacent ones do. Thus

firing pattern A will be effective, whereas firing pattern B will not:



PATTERN A PATTERN B

It is evident that we can impose "random" requirements upon each individual element, such that it fires exactly half the time and that its firing probability on any occasion is invariant with respect to its preceding firing series, and still be free to select *patterns* which will yield anywhere from zero to 100 per cent activation of the controlee system. A statistical test will tell us whether there is evidence of a "pattern-selection bias," which if found would be evidence for the operation of a "systematic" selector agent, e.g., a psychoid. (For extended discussion see Meehl et al., 1958, Appendix E.) It is assumed, of course, that the received physical$_2$ nomologicals provide no explanation for the tendency. Whether the (V$\rightarrow$ Ψ) correlation is stochastic or nomological depends upon the level of causal analysis. At the molar level, our autocerebroscopic data discorroborate any (V $\twoheadrightarrow$ Ψ) nomological, because they show that a red-type cortical state *r* is sometimes coexistent with phenomenal green. However, since the Φ-Ψ relation is many-one when Φ is microcharacterized, it is logically possible for the aberrated Ψ-occasions to be either nomological or stochastic functions of the V-states when these are subdivided on the basis of their microproperties.

I do not wish to defend such a psychoid theory, which is admittedly rather impoverished in content (although, I think, not empty or frivolous). On the evidence stated, it would seem imparsimonious to postulate such enduring continuants as psychoids. We might better conceive of some sort of "Ψ-field," an occurrent characterized by suitable phenomenal quasi-functors, and acting back upon the physical$_2$ system. Examples like self-induction in physics are helpful in dispelling the anxiety sometimes aroused by the "emergent" features of such psychophysical theories. Of course self-induction is emergence only epistemologically, not ontologically, since Omniscient Jones knows that the field about a coil—its existence and all its quantitative features—is derivable from statements about the fields associated with the elementary particles of which current-in-coil is composed. A scientist who *had* already carried theoretical analysis of electric current to a "deep" enough microlevel would have been able to *predict* the self-induction effect. So we have here emergence in the context of discovery, but not emergence in the "ontological" sense required by the physical$_1$-physical$_2$ distinction. This clarification may itself be helpful in getting clearer about the physical$_1$-physical$_2$ distinction. How could we know whether to call a theoretical entity like a "Ψ-field" physical$_1$ or physical$_2$?

The Meehl-Sellars definition of physical$_2$ has the oddity of making a *world-historical* reference in stating the nature of a theoretical construct. "An event or entity is *physical*$_2$ if it is definable in terms of theoretical primitives adequate to describe completely the actual states though not necessarily the potentialities of the universe before the appearance of life (Meehl & Sellars, 1956, p. 252). Suppose Utopian neurophysiology, even after discovery of the

aberrant autocerebroscopic effects, is able to provide an adequate causal account of everything that happens by speaking of nothing but elementary particles, electromagnetic and electrostatic fields, etc. The relation of raw feels to "life" might be explicable in terms of certain structural peculiarities of complex carbon molecules, such that the configural relations among elementary physical$_2$ functors needed to render any raw-feel quasi-functor nonzero cannot arise without a kind and order of complexity to which only the carbon atom lends itself chemically. The raw-feel quasi-functor is then a number characterizing certain mathematical features of the physical$_2$ functor *relations*, which is why it is permissible to speak of "identity." That is, the raw-feel event *r* is physically *constituted* or *composed* of the elementary particles with their associated fields and forces. A raw feel is an occurrent, whose physical$_2$ nature is a certain configuration of elementary physical$_2$ continuants; and the "richness" of this configuration presupposes a structural and causal complexity not physical$_2$-possible except with carbon compounds. If it should be technically possible to synthesize organisms built around silicon, such androids (Scriven, 1953) would be confidently said to "have experiences," and of course their verbal and other expressive-reportive behavior would be consistent with this. The notion that they might "merely be talking *as if* they 'had experiences' " would not be a theoretically admissible notion, since we constructed them and "a raw feel" would be *constitutively* defined by reference to the configuration of fundamental physical$_2$ functors. Similarly, such classical puzzles as the experiences of dogs or earthworms would be soluble, simply by substituting in the expressions for phenomenal quasi-functors the physical$_2$ functors determined experimentally by studying the brain functions of these creatures.

How could there be any such entity as a physical$_1$ raw-feel event or state which was not physical$_2$ by the Meehl-Sellars criterion? It would have to be an entity not constituted of physical$_1$ entities, an entity not composed of physical$_2$ "phases" or "parts" or "substates," and which literally *comes into being* for the first time when physical$_2$ entities enter into a certain configuration. The occurrence of this requisite physical$_2$ configuration is necessary and sufficient for the existence of the new entity, but the latter is not itself the configuration. It seems that we would have here a truly fundamental nomological, a rock-bottom feature of the world. This brute fact would qualify as one of the "insoluble mysteries of the universe," and Du Bois-Reymond would have been vindicated in calling it one of the seven such. I do not know whether there are any comparable cases in current physical$_2$ science; but it is safe to say—as a matter of logic!—that at *any* given stage of knowledge, even the fictitious "final, Utopian, complete" stage, there will necessarily be some primitive theoretical propositions. One or more of these may be purely dispositional, designating properties whose actualization is contingent within a world-family and "novel" at time *t* in an actual world. Such a property would be a true emergent in one acceptable sense of that tricky word. Professor Paul Feyerabend has suggested to me the physical$_2$ possibility of a two-particle universe, say a pair of electrons, placed at a distance such that their gravitational attractive force exactly equals their electrostatic repellent force. Stable thus, no electromagnetic field exists; but this universe has the potential to develop one if somehow (e.g., by the finger of God) a relative motion were to occur between them.

These considerations show that a non-physical$_2$ Ψ-field might be similar to what we ordinarily call "mental," or it might not. Suppose all efforts to analyze it failed, and we found it impossible to develop a theory of its fine structure, to break it conceptually into parts or components, to specify any sort of spatial regions or intensity gradients—in short, to say any-

thing about it except its causal role in the physical$_1$ brain-system. Now this would be a very exciting scientific position, because our analysis of case (2a) showed a kind of "teleology" in the selective influence exerted by Ψ over the locally indeterminate physical$_2$ outcomes in the tokening mechanism. Crudely put, we could say that "Whatever the nature of Ψ may be, as a causal agent it 'acts purposefully,' it throws the physical$_2$-indeterminate subset of micro-events in system T 'so that' the molar out come is the meaningful tokening t_r, rather than the physical$_2$-probable disjunction of micro-outcomes which would lead to a meaningless pseudo-token (e.g., a neologism), or a jamming of the mechanism, or tokening something inappropriate to the modality (e.g., 'C sharp' when the current physical input is from the visual cortex). All efforts to microanalyze Ψ having thus far failed, all we are able to say about it is that it is a non-extended, homogeneous, unitary, non-space-filling whoozis, which mediates tokenings by teleological selection of subsets of physical$_2$ micro-events." I suggest that it would be quite appropriate to say that such a combination of negative and positive properties and powers is rather like the "mental" of traditional dualism. If I am right in this, it means that a radical ontological dualism must be regarded as having empirical—even "scientific"—meaning, contrary to what was alleged by the Vienna positivists and some of their philosophic descendants.

A philosophically relevant result of this analysis of the scientifically available outcomes of our thought-experiment is that the admissibility of "I seem to be experiencing green but I am really not because my cortex is in state *r*" hinges upon prior acceptance of a specified nomological network. If case (1) obtains, no one would ever be impelled to say such a thing, but it would be admissible because a person *could* (physical$_2$ possibility) have a structurally defective tokening mechanism. If case (2a) obtains, the paradoxical remark might be made, and correctly. If case (2b) obtains the remark would be proper or not depending upon whether the net allows for "slippage" between Ψ and T, between *r* and Ψ, or both. The micronomologicals between *r* and Ψ (i.e., laws relating phenomenal quasi-functors characterizing Ψ to their determining physical$_2$ functors in *r*) might be such that "slippage" between the molar *r*-state and phenomenal qualities is theoretically derivable, whereas the psychophysical correlations relating Ψ and *r* jointly to the molar tokening events are strict (nomologicals). If that were so, an observer would know (scientifically) that any impulse to token the paradoxical sentence should be resisted, because the object-language "I am experiencing green" will always be correct. (We assume here an autocerebroscopically confirmed *ceteris paribus* regarding other potentially interfering cerebral systems, such as Freudian slips.) Per contra, if the total evidence corroborates a network in which the ($r \nrightarrow \Psi$) law is tight and the joint ($\Psi \bullet r \nrightarrow_p t_r$) law loose, the paradoxical statement is not only admissible, but mandatory. Intermediate cases (both anchorings of Ψ stochastic) would lead to varying probabilities, the paradoxical statement being sometimes right and sometimes wrong.

III. Some Alleged Metapredicates of Raw-Feel Statements

In the light of this thought-experiment, let us examine some of the metapredicates traditionally attributed to raw-feel statements, together with some of the familiar grounds for attributing them. I shall distinguish (without prejudging their independence) claims that raw-feel statements are (not perhaps always, but sometimes) *noninferential*, *incorrigible*, *inerrant*, *indubitable*, *private*, and *ineffable*.

1. *Noninferentiality*: Lord Russell, precisely reversing the line of methodological behaviorism, distinguished the physical from the mental by the epistemic criterion that the former is inferred and the latter is not. Both Russell and the (neo)behaviorists are right, inasmuch as the mental events of other people are inferred by me but noninferred by them; and my mental events are inferred by other people but noninferred by me. This is true only if "mental" is taken as synonymous with "phenomenal," a convention which is so inconvenient for clinical psychology that it has been abandoned there. But in the present context, where the mind-body problem is stated in terms of its *raw-feel component* rather than its *intentionality component*, Freud's theories are irrelevant. Our "mental" is—roughly—Freud's "conscious." The intentionality component of the mind-body problem has been solved, in its essentials, by Sellars (1953). While expressions such as "immediately given," "known by acquaintance," "hard data," "true by ostensive definition," and the like have been powerfully criticized as misleading, equivocal, and downright false, still it is generally agreed that these expressions all aim at *something* which is uniquely true of raw-feel statements. Just what that something is remains a matter of controversy, and I have here chosen "noninferential" as the least misleading, least disputed, and most central or "core" component of the explicandum.

It is important that "noninferential" does not mean *noninferable*. If raw-feel events are in the physical$_1$ network, then raw-feel statements are inferable from non-raw-feel statements. I presuppose throughout that phenomenal events, whether physical$_2$ or not, are physical$_1$. If phenomenal events were not even stochastically *correlated* with human speech, writing, or other signals, nor with stimulus events, nor with internal bodily (brain) events, there would be no "mind-body" problem, no identity theory, and no conversation about such matters among philosophers. It is doubtful whether one could even be said to "know about" his own phenomenal events in such a world (see Meehl, 1950). If Jones tokens "I see red," Smith may infer probabilistically that Jones sees red; Smith's evidence for "Jones sees red" is the statement "Jones tokens 'I see red,' " a statement whose *content* is behavioral (and linguistic), not phenomenal (and nonlinguistic). Smith may alternatively infer "Jones sees red" from other statements, such as "The apparatus is transmitting 7000 Å light waves to Jones's retina," "Jones is fixating a neutral gray wall after having looked for a minute at a green circle," "The neurosurgeon is electrically stimulating Jones's brain (etc.)," "Jones's GSR changes in response to stimulus word 'red.' " These inferences are probabilistic, but they are (probabilistically) valid inferences. Since these other evidential statements are also available to Jones, it follows that he *could* infer his own phenomenal statement, as Smith does. So we see that the autocerebroscopic situation is not epistemologically unique, but merely tightens the net by providing more "direct" readings from the brain.

We have now put the necessary hedges around the claim that raw-feel statements are noninferential. It does not mean that they *cannot* be inferred; it does not mean that they are never *in fact* inferred; it only mean; that they are *sometimes made without being inferred*, and typically so by the knower whose raw feels are their subject matter. The familiar sign that a raw-feel statement belongs to the class of noninferred raw-feel statements is, of course, the statement's use of egocentric particulars Jones *may* properly say "Jones sees red," and in philosophical or autocerebroscopic contexts he might actually adopt the third-person locution but normally Jones says "I see red" to designate that state of affairs which Smith would designate by "Jones sees red."

Although other statements *might* be adduced as evidence for "I see red," and although

I may now or later admit evidence against "I see red," it is nevertheless true that knowers sometimes token "I see red" without having antecedently tokened any statements from which "I see red" could be inferred. This fact of descriptive pragmatics is perhaps the minimum content of a claim that raw-feel statements are noninferential; but what is philosophically relevant is our generally accepted belief that such noninferential tokenings are sometimes legitimate moves, so that a knower who tokens "I see red" without antecedently tokening any statements from which it can be inferred is not necessarily tokening illegitimately or irresponsibly. How can such a thing be?

It has sometimes been said that the "evidence" which justifies a raw feel statement (under the usual noninferential conditions) is "the experience itself." I do not wish to condemn such talk as utterly without merit, since I believe that it intends something true and fundamental. But if we adopt the generally accepted convention that "evidence for..." means "providing a basis for inferring..." and remind ourselves that inferability is a relation between statements or propositions, it follows that we cannot properly speak of an experience (something nonpropositional) as being "evidence for" a statement (something propositional). It seems then that noninferential raw-feel tokenings are, strictly speaking, tokenings of statements "in the absence of evidence." On the other hand, we do not ordinarily countenance the tokening of raw-feel statements by a knower who is not concurrently experiencing the raw feel designated by the statement tokened. If Jones makes a practice of tokening "I see red" on occasions when he is not in fact seeing red (a fact which we infer from other evidence, which may include his own subsequent admission that he lied, or mis-spoke, or "just felt like saying it") we consider him irresponsible, because he tends to token illegitimately.

I do not have anything illuminating to add to what others (e.g., Sellars, 1954) have said by way of expounding what is involved here. A knower is considered to token egocentric raw-feel statements legitimately when his tokening behavior is rule-regulated, whether or not he antecedently tokens the rule itself (which he normally does not). There is a degenerate, uninformative semantic rule, " 'Red' means *red*," to which an English-speaking knower's tokenings may or may not conform. It may be viewed, within pure pragmatics, as a language-entry rule, conformity to which legitimates an egocentric raw-feel tokening. We may employ special words like "legitimate" (applied to particular tokenings) and "responsible" (applied to a knower who tokens legitimately) as distinct from the word "rational," since the latter refers to intertokening (intralinguistic) relations, e.g., inferability, conformance to language-transition rules, which we have seen is not at issue for the typical egocentric raw-feel statement. A tokening "I see red" by a knower who is not concurrently the locus of a red-qualitied raw-feel event is false; the act of so tokening is illegitimate (i.e., semantic-rule-violative); knowers who tend to token illegitimately are sense-defective, psychotic, or irresponsible. We need terms other than "irrational" in this context, because we reserve the latter for a disposition to violate language-transition rules, i.e., to token against, or without, (propositional) *evidence*.

Speaking epistemologically, the legitimacy of tokening egocentric raw-feel statements without antecedently tokening any statements as evidence (and in most cases, without being able to do so upon demand) depends upon the tokening's conformity to the semantic rule. There is in this notion of a legitimate egocentric tokening a small but inescapable element of "psychologism," that amount and kind of reference to the nonlinguistic which is not a philosophic sin because it falls under the heading of pure pragmatics—the *coex*

relationship of Sellars (1947, 1949). Speaking psychologically, the English-speaker's overwhelmingly strong impulse to token "red" when experiencing a red-qualitied raw feel has its causal account in intimate wiring plus verbal reinforcement history. We are all trained against tokening non-raw-feel statements without antecedently tokening statements from which they can be inferred, or at least without being able to offer such on demand. "Just because," or "I say what I mean," or "I need no evidence" are childish or neurotic responses to "Why do you say that?" for non-raw-feel statements, and are finally beaten out of all but the most irrational. But we are permitted to token raw-feel statements without evidence, so long as we do it in a rule-regulated way (i.e., in what the reinforcing verbal community takes, on all of *its* evidence, to be conformable to the language-entry rule " '*Q*' means *Q*").

2. *Incorrigibility*: It has been shown above that noninferentiality does not mean noninferability, that while egocentric raw-feel statements are commonly and legitimately tokened without inference, their allocentric equivalents are regularly inferred and even the egocentric ones can sometimes be. Since raw-feel statements can be inferred, it follows directly that their contraries can be discorroborated. Hence a raw-feel statement "Jones sees green" can be argued against by evidence supporting "Jones sees red," given the grammar of these phenomenal color predicates. And while this is clearest in the case of allocentric raw-feel statements, it is true of egocentric ones also. Since evidence supporting "Jones sees red" is evidence against "Jones sees green," and such evidence is also, at times, available to Jones himself, it follows that Jones can have evidence tending to discorroborate his egocentric raw-feel statements. It is puzzling why so many philosophers have doubted this, when it follows rigorously from two universally accepted theses: (a) Raw-feel events have knowable causes and effects (either would do!). (b) Some raw-feel statements are incompatible with others.

Incorrigibility of a raw-feel statement would require that, while counterevidence can unquestionably be brought to bear upon it, the amount and character of such counterevidence could never be sufficient to warrant abandoning it. While this doctrine seems widely held, I have never seen any cogent proof of it. I personally find the autocerebroscopic thought-experiment very helpful in this respect. My dilemma as Utopian neurophysiologist is to decide whether any sufficiently massive and interlocking evidence could corroborate "I am seeing red" more than all of my evidence corroborates the claim that I always token in obedience to the language-entry rule " 'Green' means *green*." It cuts no ice at all to "feel sure," since "Meehl feels sure that he always tokens in conformity with the language-entry rules" clearly does not entail that he does in fact so token. The kind of evidence hypothesized in our thought-experiment provides an adequate explanation of my aberrated "green" tokenings, as well as of my strong feeling of subjective certainty. Given that other evidence can be *relevant* to the truth of a raw-feel statement, it is hard to see any reason for insisting that it could not conceivably be countervailing against the unevidenced raw-feel statement, unless we accept the metaproposition "All egocentric raw-feel statements are true," a proposition which is surely not analytic. And, if it is itself empirical, it cannot be made untouchable by all amounts of discorroboration.

3. *Inerrancy*: Without making the strong claim of incorrigibility (no egocentric raw-feel statement could, *in principle*, be opposed by evidence sufficient to require its abandonment), it might be hoped that nevertheless no such statements are *in fact* false. A specious plausibility is lent to this view by the widespread habit of putting "et cetera" after a list of interfering factors which are at present known to produce illegitimate tokenings. Thus a

philosopher writes, "An English-speaking plain man, not lying or afflicted with aphasia, etc., cannot err in tokening 'I see red.' " Great danger lies in the innocent-looking "etc.," which has to do duty for all the unlisted sources of mistokening. The statement that one *cannot* token an egocentric raw-feel statement incorrectly without misunderstanding English, being aphasic, having a Freudian slip, being drugged, obeying a posthypnotic suggestion, *and so forth* is analytic only if "and so forth" is short for "any other causal source of mistoken–ing." Can the list be written out, substituting an extensional specification for "etc.," rather than the intensional "causal source of mistokening"? Obviously no such extensional list could be compiled, lacking an admittedly complete causal theory of experiencing-and-tokening. Is anyone prepared to say on present evidence that we already *know* all possible sources of mistokening? *Could* anyone claim this while the mind-body problem remains *sub judice?* If anyone did so claim, would his claim be incorrigible with respect to the auto–cerebroscopic experiment? What this argument for inerrancy really comes down to when explicated is a tautology plus a bow to psychic determinism, namely: "An egocentric raw-feel statement cannot be erroneously tokened unless some disturbing factor leads to knower's violating the language-entry rule," i.e., an egocentric raw-feel tokening is legitimate unless something causes it to be made illegitimately. This triviality cannot sustain a philosophical claim of inerrancy. Since we already have indisputable evidence that mistokenings do in fact occur, why would we venture on the unsupported empirical claim that the presently known list is exhaustive?

Neither common-sense examples nor philosophical arguments suffice to destroy in some people the conviction that "I just *couldn't* be wrong in 'attentively (etc.)' characterizing the hue-quality of my momentary raw-feel." By way of softening up such a reader, I recommend study of a fascinating experiment by Howells (1944) in which he showed that repeated association between a color (red or green) and a tone (low or high, respectively) produced in his subjects a growing disposition to mislabel the colors when presented unsaturated along with the "inappropriate" tone. The experimenter infers (as I do, tentatively) that the subjects came to actually *experience* the wrong color; other psychologists opt for the alternative hypothesis that they came to *token* erroneously. It seems rather implausible that VIIIth-nerve inputs to Heschl's gyrus could acquire such control over the visual-experiencing cortex; but it also seems hard to believe that they began to say "red" when seeing green. The point is that Howells's results are compatible with either interpretation; and reading them should, I think, at least attenuate one's intuitive faith in the "obvious infallibility" of raw-feel statements.[3]

4. *Indubitability*: By this word I mean the psychological inability to doubt our egocentric raw-feel statements at the time we token them. Does this inability, if it in fact obtains, have any philosophic relevance? I think it has none, unless we presuppose that "Jones is psychologically incapable of genuinely doubting *p* at time *t"* entails, nomologically or analytically, "*p* is true." I take it that no one will wish to maintain that a knower's psychological lack of power to doubt a proposition implies the latter's truth, the implication being either a law of nature or a truth of logic. I do not myself believe that psychological indubitability is even incompatible with the knower's sincere assertion that *p* is corrigible. The reader will have to judge this for himself, with respect to "sincerity of assertion." But the metalinguistic statement "I know that there could be strong countervailing evidence against *p*" is surely compatible with "As of this moment, I find myself unable to entertain seriously the notion

[3] I am indebted to my colleague Dr. Milton Trapold for calling this experiment to my attention.

the *p* is in fact false." In the autocerebroscopic situation, our Utopian neurophysiologist continued to collect discorroborative evidence against his own raw-feel statements, because he rationally recognized that he *ought*, if such counterevidence were forthcoming, to correct these raw-feel statements. But during the course of this evidence-collecting, he "felt certain" that a bug would be found in the apparatus, because the raw-feel statements were currently indubitable by him. This would be "inconsistent" cognitive activity for one who held that a knower's inability to doubt entails the truth—analytically, in logic or epistemology—of the indubitable statement. But our neurophysiologist holds no such indefensible view.

5. *Privacy* ("privileged access"): It is agreed that no other person is the locus of my raw-feel events. This simple truth can be formulated either epistemically or physiologically, as follows: (a) A raw-feel event x which belongs to the class C_1 of events constituting the experiential history of a knower K_1 does not belong to the class C_2 of a different knower K_2. (b) The tokening mechanism whose tokenings characterize the raw-feel events of organism K_1 is wired "directly" to K_1's visual cortex, whereas the tokening mechanism of K_2 is not directly wired to the visual cortex of K_1.

It seems that (a), stated in philosophic language, is an analytic truth (but see below), whereas (b), stated in physiological language (except for "tokening"), is synthetic. Query whether the contradictory of the second half of (b), while counterfactual for all historical organism-pairs, is counternomological? Given Utopian neurophysiology, we can at least conceive a procedure whereby the tokening mechanism of K_2 might be wired to the visual cortex of K_1 in such a way that the "causal intimacy" of K_2's tokening events would be as "close" to the events occurring in K_1's visual cortex as are the tokening events of K_1. Or, without such direct anatomical wiring, Utopian neurophysiology could presumably impose a pattern upon K_2's tokening mechanism which was physical_2-indistinguishable (within tolerance limits set narrower than those yielding "no difference" reports or discriminations for the single specimen) from the pattern imposed by K_1's visual cortex upon K_1's tokening mechanism. While I don't wish to liquidate any valid philosophical problem by shifting the grounds to biology, it is worth emphasizing that from the standpoint of physical_2 causality, the "privacy" of my raw-feel events consists of the rather unexciting fact that "I" (= the person whose body is the locus of the tokening mechanism which tokens the egocentric raw-feel statement in question) am causally related to "my" (= visual-cortical states located within that body) raw-feel events in a close-knit, causally direct way; whereas "you" (= the person whose body is the locus of the tokening mechanism which tokens the allocentric equivalent of the egocentric raw-feel statement in question) are causally related to "my" raw-feel events by a causal chain having more links and therefore, normally, having greater possibility of "slippage" (i.e., of intervening nuisance variables which may prevent complete correspondence). From the physical_2 point of view, there is nothing mysterious or special about the fact that Jones knows he has a stomach-ache under circumstances when Smith does not know that Jones has a stomach-ache. Jones's knowing (= tokening) system is linked fairly directly to Jones's stomach; Smith's knowing system is normally linked to Jones's stomach only via Jones's knowing system, thus: Jones's stomach → Jones's tokening system → Jones's reporting system → Smith's receptor system → Smith's tokening system. It being a physical_2 fact that each of these causal connections is only stochastic, every link involves a further attrition of probabilities: hence the correlation between events in Smith's tokening system and events in Jones's stomach will be lower than that between events in Jones's tokening system and events in Jones's stomach.

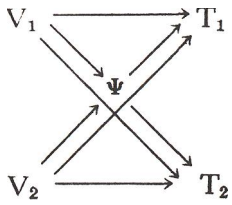
Such an analysis of the causal basis of the epistemic situations suggests that, in principle, there *could* be special physical$_2$ arrangements in which the linkage between Smith's tokening system and Jones's stomach would be more dependable than that between Jones's tokening system and the latter's own stomach. And we already know of such clinical examples, e.g., the recently operated patient who "feels sensations in his leg" when the surgeon knows that the leg has been amputated. Such examples serve to remind us that the "privacy" which is philosophically interesting involves central (brain) events rather than peripheral conditions. The autocerebroscopic experiment pushes this "privacy" as far back as it can be taken from the physical$_2$ standpoint, and this extreme is the *only* stage of causal analysis that is of intrinsic philosophical interest for the mind-body theorizer. Here we have the situation in which two quasi-nomological links are set in (apparent) opposition. The subject's tokening system is considered to be very intimately linked to his visual raw-feel states; and the autocerebroscope is also very intimately linked with them. If the Utopian microanalysis supports Feigl's identity theory, either in form (1) or (2a) above, we must conclude that the usual "privacy" (= privileged access) of raw-feel events merely reflects the usual organic condition of intimate wiring, and the usual lack of an equally tight independent access; but that special physical$_2$ conditions *can* be set up in which the causal linkage between artificial apparatus and raw-feel events is more trustworthy than that provided by nature's cerebral anatomy. The progress of science will merely have carried the "stomach-to-brain" kind of analysis inward, so to speak; and educated Utopians will recognize that scientific instruments are (while fallible) less fallible than the tokening mechanism with its socially learned linkage to events in other cerebral subsystems. Less educated Utopians will absorb this understanding by downward seepage through the Sunday supplement section, and the Utopian "plain man" will feel nothing absurd about such locutions as "Jones thinks he is experiencing a pain, but he's really not."

Given the frame of Utopian identity theory, analytic thesis (a) can also be *stated* physicalistically, but it is no longer clearly analytic. We say, "A raw-feel event occurring in the visual cortex of organism K_1 cannot occur in the visual cortex of organism K_2." (I mean this literally, of course; there is nothing analytic, or even plausible, about a denial that raw-feel events having the *same properties* can occur in two different brains. The statement is about identity of particulars.) This would be frame-analytic given a biological law "A visual cortex V cannot belong to two organisms K_1 and K_2 at the same time," which is contingently true but not clearly nomological, since Utopian neurophysiology may provide surgical techniques whereby two organisms can share their visual cortices. Admittedly such a Siamese-twin brain-fusion would necessitate adoption of a convention as to what "two organisms" will be taken to mean. But it would seem to be a reasonable, nonarbitrary stipulation that if two bodies were so joined as to possess a common visual cortex but shared nothing else (e.g., each has a separate tokening mechanism wired equivalently to the common visual cortex; each has its own motivational cortical subsystems; each retains its own stock of memory-storage and "self-concept" subsystems), Utopian physiology would properly speak of "two organisms having a shared visual cortex." The raw-feel *knowings* of these joined individuals being conventionally taken as their distinct propositional (= tokening) events, Utopian epistemology would allow the locution "Knowers K_1 and K_2 share visual raw-feel events," so that (a) not only is seen to be nonanalytic but is empirically false. It is worth noting that this Utopian situation is not utterly unlike our present situation, when we say that "Raw feel x belongs to knower K," even though knower K may be identifiable over a

twenty-year period only by the fact of genidentity (i.e., most of the matter constituting K's brain has been replaced during the twenty-year time interval). It is not easy to say precisely why a particular twenty-years-past raw-feel event belongs to "K's set" of raw-feel events *except* that it belongs to the set historically associated with a body genidentical with K's present body. This line of thought may be worth pursuing on another occasion, because it suggests the possibility that the inaccessibility of another's mind amounts fundamentally to the same problem as the reliability of memory in relation to the solipsism of the present moment; so that he who "resolves" the latter—by whatever means—will thereby have also resolved the former.

How would the privacy problem shape up if Utopian neurophysiology were led by the evidence to adopt an ontological dualism of the kind hinted at in Section II? Surprisingly enough, this does not seem to make any important difference. It having been shown experimentally that (non-physical$_2$) raw-feel events are causal descendants of physical$_2$ brain-events in the visual cortex and that a raw-feel event produced by a state of cortex V_i cannot influence tokening events in tokening system T_i unless T_i is simultaneously "physical$_2$-close" (and perhaps neurologically wired) to V_i, we adopt the location-convention that a (non-physical$_2$) raw-feel event "occurs at (or, even, *in*)" the brain B which contains properly juxtaposed and wired subsystems V_i and T_i and in which the nomological subnet

is being instantiated. The physical$_2$ conditions necessary and sufficient for such instantiation would be known, and it would be a technological problem to realize them by surgical methods of brain-fusion. Suppose that somehow, without hopelessly disrupting the general cerebral arrangements needed for preserving minimal "personhood," we could juxtapose each of the two tokening systems T_1 and T_2 to each of two visual systems V_1 and V_2, so that the surgeon's electrical stimulation of *either* V_1 or V_2 would produce raw-feel events which influence events in *both* T_1 and T_2. Our experimental analysis continues to support the notion that only a single raw-feel event (or short-lived continuant) is produced by such stimulation. Our dualistic network looks like this:



When we stimulate V_1, both subjects token "I see red." Similarly they both token "I see red" when we stimulate V_2. There seems to be no reason, behavioristically, introspectively, or neurophysiologically, for attributing either of these raw-feel events to one of the subjects rather than the other, unless we adopt a convention that the person whose *visual* cortex is causative of (an otherwise shared) raw-feel event is the person who "has" it. Given a Utopian psychophysiology *in dualist form*, such a convention would, I think, be rejected as arbitrary, counterintuitive, and systemically inconvenient. Even adopting such a convention, however, we can modify the situation by reverting to a shared cortex—a third person's, or a chimpanzee's, or even a synthetic one!—-and dealing with a setup diagramed thus:

Here everything is symmetrical. The "red" physical$_2$ state Φ_s occurring in (shared) visual cortex V_s produces one raw-feel event Ψ_s (= phenomenal red) which in turn combines with the physical$_2$ neurophysiological links to elicit appropriate tokenings in both T_1 and T_2. Surely we do not want to say, "This raw-feel event occurs, but belongs to no person's experiential history." Unless the now deceased chimpanzee is a candidate, the raw-feel event must be assigned to one of our subjects; and there is not the slightest reason for choosing between them. Note that physicalism and methodological behaviorism are not involved here. The ontology is explicitly dualistic; and if you were either subject, you would indignantly oppose a philosopher's asserting that—because of the oddity of the context—your momentary raw feel "does not really belong to you, but belongs to the other fellow (or the deceased chimpanzee)."

The most fascinating feature of these fantastic thought-experiments and the feature having most philosophic relevance, is that they force us to re-examine the analyticity of epistemic-pragmatic statements about "I," by bringing out our implicit assumptions concerning the nature and identity of *persons*. Unless a person is a "simple"—an entity without "parts," "regions," "components," "subsystems"—there must be some criteria by which part-events are assigned to their appropriate persons. In phenomenal or dualistic language, we currently take it as analytic that "a particular raw-feel event cannot belong to the class C_1 of raw-feel events constituting the *Erlebnis*-history of person I and also to the class C_j constituting the *Erlebnis*-history of a nonidentical person J." But the analyticity of this is not a simple dictionary-analyticity like "No woman x is the wife of any bachelor y." If analytic, it is frame-analytic, involving the nomological network which embodies our received metaphysic of "persons." If a "person" is *defined* by specifying a class of raw-feel events, how is the class to be specified except by reference to the genidentity of a body whose cortical states are the immediate causal ancestors of the raw-feels being classified? And if we do it this way, the statement about persons' raw-feel sets being disjoint hinges upon the supposition that two bodies' cortical states must necessarily be disjoint. This in turn depends upon the alleged impossibility of two brains sharing parts of cortices—namely, the "parts" whose states are the immediate causal ancestors of raw-feel events. I cannot see any clear reason why such a state of affairs is impossible.

One must, of course, beware of the temptation to imagine an inner observer-of-raw-feels, a manikin who is located elsewhere than the visual cortex and who "inspects" the raw feels occurring in the visual cortex. The locus of the phenomenal red quality is the raw-feel event Ψ, and the only "manikin" is the tokening mechanism, which *tokens* but does not *experience*. If someone insists that by "I" he means "whatever subsystem is the raw-feel locus," then there are not two "I"-persons when the brains are conjoined so as to share a visual cortex, but one "I." We have then to remind such a subject that he must avoid token-reflexive discourse about his visual raw feels, because the "I" who *tokens* is not the "I" whose raw feel is being tokened about. Alternatively, if the subject insists that token-reflexive statements be allowed because, he holds, the "I" referred to is not genidentical with that psychophysical system whose tokening mechanism is being used but rather with that

psychophysical system which contains the visual cortex genidentical with the experiencing one, then he is committed to saying "I am a (dead) chimpanzee's brain."

The upshot of these considerations seems to be roughly this: If "privacy" means "privileged access," it is a matter of degree as regards the closeness (tightness) of causal connection between a tokening mechanism and a visual raw-feel event. Under ordinary circumstances (lacking autocerebroscopes, supersurgery, and the like) the causal chain connecting a person's raw-feel events to his tokening events is more trustworthy than the chain connecting his raw-feel events to another's tokening events. But this is no absolute principle, for the same reason that a person might sometimes be well advised to accept the autocerebroscopic verdict on his *own* raw-feel events in preference to the deliverances of his own "direct-line" tokening mechanism.

If "privacy" is taken to mean "absolute noninspectability," in the sense that the non-overlap of two persons' raw-feel-classes is analytic, I have tried to show that it is at least not dictionary-analytic but, at most, frame-analytic within an (arguable) nomological network about "persons." And I have *suggested* that this network of constraints about what is possible of persons may really not be nomological but only contingently universal, and perhaps purely technological.

How stands it with the old puzzle about whether my raw-feel qualities are qualitatively the same as yours? If the identity thesis is correct, the problem is solved experimentally, by investigating the physical$_2$ functor conditions present in your brain when you token "red" and in my brain when I token "red." (I hold, as against Professor Feigl, that *if* his identity thesis is correct, he cannot formulate the inverted-spectrum problem in his Utopian theoretical language. It would make no more sense than it would to ask, "Granting the kinetic theory of heat, is the 'hotness' of this bucket of hot soup the same as the 'hotness' of that one, given that their molecules have the same mean kinetic energy?")

What if Utopian psychophysiology is dualistic, along the lines sketched above? We know that English-speaking subjects normally satisfy the causal law network

in which the *form* of the tokening event t_r is learned but the other connections are nomologically necessitated by the (innate) wiring diagram plus the mysterious fundamental nomologicals which entail the raw-feel event Ψ_r whenever the physical$_2$-"red" cortical state Φ_r occurs in a normally constructed visual cortex. The question is whether the raw-feel event Φ could play this causal role in two brains but be of red quality in one and green quality in another. Concentrating for the moment on the "input" side, a Utopian scientist would probably argue that some such overarching principle as "same cause, same effect" should apply here as elsewhere in scientific thinking. The intersubjective nomological network being well corroborated, we "believe in it" until further notice. It tells us that there are two events Φ_i and Φ_j which occur in the brains of persons I and J, and they have certain common physical$_2$ functor properties which are causative of two events Ψ_i and Ψ_j that also share all known causal properties. If there is a property that is nonefficacious, in the sense that it is only a causal descendant but not a causal ancestor (i.e., a nomological "dangler"), our Utopian scientist would probably feel it imparsimonious to hypothesize that this nonefficacious property was different in the two individuals. He might say, "I find no reason why it

should differ, being consequent upon similar causal conditions, so I shall assume that it *doesn't*." I must confess that this argument strikes me as convincing, but I don't know what to say if someone challenges the "same cause, same effect" principle.

Perhaps the Utopian thought-experiment can help us gain further insight into this idea of the "phenomenal red quality" as a noncausal property of the raw-feel event Ψ_r. If the phenomenal red quality is noncausal, what can the privileged-access dualist be taken to mean by raising his question? If I am a dualist arguing for absolute, analytic privileged access, how do I frame the question in Utopian physiologese? I say, "While scientific findings may show that my raw feel Ψ_r has causal properties identical with your raw feel Ψ_r, I cannot infer—even in probability—that the *non*causal property *red quality* is present in yours, whereas I know that it is present in mine." Since we have agreed that merely experiencing a raw feel cannot be identified with "knowing that it has a certain quality," I as a dualist must be at least claiming that I can token "red" in obedience to the language-entry rule " 'Red' means *red*." (I do not raise here the incorrigibility or infallibility question, settled above in the negative. All that I claim for the sake of argument is that I "know" in the limited sense of high-confidence probability.) Is my tokening "red" in any way determined by the presence of the red quality in Ψ_r? Not if this quality is taken to be noncausal. If this is true, then my tokening in obedience to the language-entry rule is solely due to the *correlation* between the (causally inefficacious) red quality and the (causally efficacious) physical$_2$ event Φ_r conjoined with the *other* (causally efficacious) properties of Ψ_r. I do not, on such a view, token correctly *because* of the red quality, but because of other factors. *If* I tend to token in obedience to the language-entry rule, it is only because of the nomological dangler which relates the noncausal red quality of Ψ_r to its antecedents. *Consequently I have no stronger ground for trusting my own raw-feel tokenings than for trusting those of others.* In both cases, I must rely upon the "same cause, same effect" principle to infer the noncausal red quality from its *input* conditions. If the phenomenal quality is conceived of as noncausal, as a dangler, I can't know—except by trusting the dangler—that you experience red qualities; but I can't know that I do either!

One way of saying this is that, although a semantic rule and a causal nexus are two very different things, yet in the case of noninferential egocentric raw-feel statements, "legitimated" solely by their obedience to the language-entry rule, *a semantic tie requires* a *causal tie.* One need not, of course, *assert* the causal tie to "justify" such tokenings as they occur. But if he repudiates the claim that there *is* a causal tie, he cannot maintain in metalinguistic discourse about "privacy" that he trusts the semantic tie and therefore "knows" that he correctly characterizes his own raw-feel events.

There is something at least pragmatically strange—I do not say literally inconsistent—about Professor Feigl's view that phenomenal qualities are not intersubjective, not part of science, not in the public domain. Let us set aside the question as to whether an intersubjective test exists for the inverted-spectrum hypothesis, and consider the broader issue whether the world of science, the causal order, the physical$_1$ network, finds raw-feel qualities theoretically dispensable. Why does Professor Feigl pose the mind-body problem in the first place? What is the "puzzle" which identity theory intends to solve? If you ask him whether there are phenomenal qualities, he says, "Of course there are—this is why we have a mind-body problem." I take this to mean that Professor Feigl would not philosophize about the mind-body problem (nor would he be able to understand the discourse of another who did so) except for his own acquaintance with raw-feel qualities. I conclude from this that

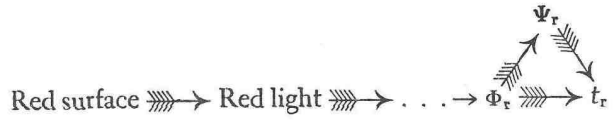
raw-feel qualities make a difference (i.e., they influence the verbal behavior of Professor Feigl *qua* philosopher). Hence they are causally efficacious, the world would be different in both physical$_1$ *and physical*$_2$ ways without them, and they must find a place in the nomological network. Must they not then be "part of science," like everything else that is causally efficacious?

It would seem so. Yet we must do justice to the claims of a rigorous and consistent epiphenomenalism, such as that defended by Lachs (1963). I am unable to detect any flaw in Lachs's incisive analysis, and feel compelled to retreat to a weaker thesis, namely, "If raw-feel qualities are dispensable from the theoretical entities of physical$_1$ science, as Professor Feigl maintains, then he must also hold that *his* concern with the mind-body problem does not originate from his own acquaintance with raw feels. Hence, he must hold that he would philosophize about the mind-body problem *exactly as he does* (compare Meehl, 1950) if, counterfactually, he experienced no raw feels of any kind whatever." I do not tax Professor Lachs with this consequence, which, I take it, would not disturb him; but I know (personal communication) that Professor Feigl finds it unacceptable. It distresses him—as it does me—to be in the very counterintuitive position of saying, "When I raise the mind-body problem, I am talking about my raw feels and their qualities; the very *meaning* of the mind-body problem involves the existence of these raw-feel qualities; a being which thinks (computes, ratiocinates, engages in rule-regulated language transitions) but lacks raw feels could not *understand* the mind-body problem. Nevertheless my raw-feel events have no causal influence upon my tokening behavior." The fascinating question whether, and how, a genuine semantic tie, " 'Red' means *red,"* could exist for a knower, lacking any causal (raw feel $\twoheadrightarrow$ tokening) tie, I shall not attempt to treat here.[4]

We therefore decide that the phenomenal red quality should be taken as causally efficacious. Hence the tokening event t_r is partially controllable by phenomenal qualities of Ψ_r. (I should perhaps remind the reader, in case "phenomenal qualities of a mental event" seems redundant, that Utopian psychophysiology will have introduced the dualistic entity Ψ_r as a theoretical construct required to make sense of the entire body of evidence, so that Ψ_r might—presumably would—have properties in addition to phenomenal ones.) In short, *whether* Ψ_r *is phenomenally red or phenomenally green makes a causal* (output) *difference*.

I have passed over for expository simplicity the fact that the form of t_r, a conventional sign, is learned, as is its connection with the Φ_r-Ψ_r complex. This psychological fact gives rise to the distressing possibility of systematic difference between the "content" of two persons' language-entry habits. Granted that "phenomenal red" and "phenomenal green" are causally differentiable, two persons could have learned to *token* pseudoappropriately if the ($\Phi \rightarrow \Psi$) links were consistently reversed between red and green. Can Utopian science find this out, given that the phenomenal hue is causally efficacious?

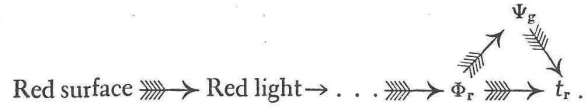
I think that it can. "Mr. Normal" has learned to token "red" according to the chain



and he therefore obeys the language-entry rule " 'Red' means *red*" consistently. "Mr.

[4] I am indebted to Professor Sellars for bringing home to me, when I was defending epiphenomenalism, the full force of this objection.

Funny" has something wrong with his idiographic ($\Phi \rightarrow \Psi$) "nomologicals" such that he experiences (consistently) phenomenal green when his visual cortex is in the red-state Φ_r. He will, however, *appear* to be normal, since he tokens in accordance with the causal system



If the language-entry rule were " 'Red' means *red-surfaced object"* Mr. Funny would be tokening obediently. And since this is the pragmatic context of verbal learning (see Skinner, 1945), he appears all right to the rest of us. When, as a sophisticated adult pursuing philosophy, he intends to shift to the phenomenal language, in which " 'Red' means *red phenomenal quality*" is the language-entry rule, he fails—but he doesn't know it (i.e., he unwittingly tokens disobediently to the rule) and neither do we. This is the inverted-spectrum problem in two colors.

According to the proponents of absolutely privileged access, this fact can never be brought to light, even by Utopian science. Let us see.

We have agreed that the phenomenal qualities must be allowed causal efficacy. That is, Mr. Funny's appropriate tokenings of "red" are causally dependent upon the copresence in his brain of physical$_2$ state Φ_r and the (consistently aberrated) phenomenal event Ψ_g. While this connection has been learned, *once learned*, *it makes a causal difference*. Suppose we perform the Utopian neurosurgery necessary to bring Mr. Funny's visual cortex into intimate causal connection with Mr. Normal's tokening mechanism, meanwhile disconnecting or functionally suppressing Mr. Normal's visual cortex. Now Mr. Normal's tokening mechanism is controllable by physical$_2$ inputs of type Φ_r and by physical$_1$ influences of type Ψ_r so as to token "red" when these are copresent. What will happen when the novel input ($\Phi_r \cdot \Psi_g$) "from Mr. Funny's visual cortex" is imposed? Without detail of Utopian microtheory, we can't say. But *something* odd should occur, since Ψ_g is causally differentiable from the usual Ψ_r, and—taken alone—should tend to produce the tokening "green" in Mr. Normal. Better yet, suppose we know from previous research that a somewhat "weakened" or "jammed" physical$_2$ input will not prevent a clear, strong raw-feel event from exerting nonchance discriminative control over a well-learned tokening process. Then we can interfere with the physical$_2$ chain between V and T while allowing the physical$_2$ event Φ_r (and, therefore, the raw-feel event Ψ_g) to occur in full strength and clearness. We might find our subject reporting, "It's a peculiar experience—nothing quite like I'm familiar with—but it *is* a color, and it's green—yet I also feel an impulse to say 'red'—but 'green' is stronger." By ingenious surgical cross-wirings, and well-timed suppressions or chain-interruptions, we could study the tokening consequences of various combinations of physical$_2$ inputs and raw-feel influences upon each subject's several brain-subsystems. The results might very well be so consistent that we would feel confident in saying, "When Mr. Funny normally tokens 'red,' his visual cortex is in the same physical$_2$ state that red light produces in Mr. Normal and the rest of us. But his raw feels are of phenomenal green quality under these circumstances. The rest of us have learned to token 'green' when our tokening mechanisms are under the intimate influence of a green-qualitied raw feel, so when *his* raw feels are allowed to control *our* tokening mechanisms, we token accordingly." But why are they Mr. Funny's raw feels? It's delightfully arbitrary to say *whose* they are. I would say that they are shared, given the Utopian supersurgery cross-wiring.

Of course I do not believe any such circumstance could arise, since I have metaphysical faith in the "same cause, same effect" principle. Another way of saying this, less offensive to modern ears, is that I decide resolutely to pursue the scientific aim of concocting a nomological net which will include everything that happens (Popper, 1962, 1959). If ontological dualism is true, we have the aim to fit our psychoids into the net (i.e., "Everything is physical$_1$, whether it is physical$_2$ or not"). The Utopian physiologist would insist that there can't be two utterly different hue qualities arising from the same brain-condition Φ_r, even though they are in different brains. And he would persist doggedly in searching for the as-yet-undiscovered physical$_2$ functor which makes Mr. Funny have the "wrong hue quality" when his visual cortex is in state Φ_r.

It may be argued against the brain-joining experiment that we *still* do not have any basis for inferring to the "private" raw-feel qualities, because it could be that "everything changes" owing to the special character of such Utopian cerebral rearrangements. This is of course logically possible. My only answer would be to point out that this possibility is always present when we invoke nomologicals to infer to states of affairs only "indirectly known." There is, so far as I can see, nothing peculiar in this respect about the mind-body problem. If I infer the chemical constitution of the stars from spectrographic evidence, and cross-check this inference by employing other avenues of inference, it still *could be* that "everything is different" out there—provided it is "different" in a sufficiently systematic way. The "privileged-access" absolutist cannot rely upon the truism "Our postulated nomologicals may not hold, in reality," unless he is willing to concede that the "privacy" of our raw-feel qualities has the same source as does the general fallibility of human knowledge. The philosophically distressing form of the "privacy" problem is not that which merely concedes the possibility of error; it is that which denies the possibility of a human knower's even getting evidence which is relevant to the question about another knower's raw-feel qualities. Since "getting relevant evidence about event Y" is a matter of latching on to other events X and Z to which Y is causally linked, absolute "privacy" means acausality. I have argued that anyone who admits that raw feels are caused has no positive reason for hypothesizing that they differ qualitatively when produced by similar brain-antecedents. And I have further argued that one who admits that raw feels are causally efficacious—in the required sense that the raw-feel quality is causally relevant as a critical determiner of our own raw-feel tokenings, making it possible for each of us to token obediently to even a purely "private" language-entry rule—must allow as logically possible a thought-experiment in which one knower can token responsively to a raw feel which also "belongs to" the experiential history of another knower. If Smith's tokening mechanism is controllable by the phenomenal qualities, making it possible for him to token consistently "red" when a raw-feel event in his brain is of red quality and "green" when it is green, then if we can locate a raw-feel event belonging to Jones's experiential history so as to give it causal control over Smith's tokenings, *what* he tokens will be evidential as to the raw-feel qualities of Jones.

It is worth noting that the inverted spectrum is usually taken by philosophers as an "obviously possible" contingency, whereas the psychological situation is actually far more complicated; and it is very doubtful whether such a concept can be made structurally consistent. Color experience involves more than hue, and hue is not independent of the other phenomenal dimensions of visual experience. When we combine the findings of phenomenologists, classical introspectionists, Gestalt psychologists, and contemporary "eclectic" students of perception, we have to deal with an extraordinarily rich, complex, interknit set of

relationships, both "internal" and "external," and we cannot treat of a simple hue dimension that is unrelated to anything else. It would probably be impossible to construct a formal model representing the visual-experience domain with hue order reversed such that the verbal reports and other discrimination behavior exactly matched that of normal-seeing persons in all respects. The same objection holds for C. I. Lewis's stronger hypothesis about interchanging sensory modalities, such that my red hue is your C sharp tone. In this "switched modality" case, things are worse than with intramodality inversions, because here even the dimensionality may not correspond. I think all psychologists would agree that no such switched modalities are possible, given the internal correlations—the phenomenal *structure*—of the several modalities as we know them. This of course does not refute a hypothesis which merely says "... some *other* qualities with isomorphic structure..." understanding that the "other qualities" cannot be any of those we know (i.e., there must be a radically new modality). With this understanding, the idea becomes rather less interesting; and its abstract possibility is still not obvious. I am not sure just what it means to say, "There *might* be a phenomenal modality different from vision, hearing, taste, touch, kinesthesis, etc.—utterly unlike any sensory qualities we know—whose structure was precisely the same as that of the visual modality." This modality must be related to inputs as vision is; it must be related to the subject's experience of his own movements in objective space as vision is (a very tricky requirement, perhaps impossible to meet); it must possess the appropriate number of phenomenal dimensions; and these must be related to one another (internally) and to light stimulation of the retina (externally) in exactly the way hue, saturation, "grain," brightness, size, distance, interposition, direction, etc., are. I am not prepared to say that there could not be such a modality; I merely wish to enter a psychologist's caveat against the too-easy assumption that it could be worked out. And the familiar form of the idea, in which a phenomenal dimension is simply reversed or two modalities are interchanged, is almost certainly impossible.

6. *Ineffability*: It is commonly said that "the quality of a raw feel cannot be communicated." As it stands, this is ambiguous. One interpretation is, "By tokening a raw-feel predicate, knower K_1 does not inform K_2 as to the quality of K_1's denoted raw feel." Whether this is correct or not depends upon the privacy issue, since if K_1 and K_2 are *in fact* tokening in obedience to the same language-entry rule, each tokening "red" when he experiences the red hue, then K_1 does communicate to K_2 by telling him, "Now I am experiencing the red hue." Of course K_1 may at times token erroneously, and on such occasions K_2 will receive misinformation. This kind of mistokening, in which K_1 fails to conform to his own language-entry rules (i.e., he tokens color words inconsistently), is not unique to the raw-feel problem but is also possible when K_1 tokens erroneously in the physical thing-language or in the theoretical language. We do not say that communication is impossible merely because particular communications may (and do) mislead. "Ineffable" means "unspeak*able*," which is a good deal stronger than "sometimes unspoken" or "sometimes misspoken." If Smith and Jones have each learned to token "red" whenever their visual cortices are in state Φ_r, or their psychoids are in state Ψ_r, then they normally communicate by so tokening. I am not here appealing to the usages of vulgar speech, i.e., what we "ordinarily take as communication," an appeal which is forestalled by the very posing of the inverted-spectrum problem. What I mean here is that if Smith and Jones do in fact obey the *same* language-entry rule, then Jones's token "I see red" communicates *fully* to Smith, and conveys to Smith no less about Jones's raw feel than Smith conveys to himself by his own raw-feel tokenings.

Another unpacking of "ineffability" is "The meaning of raw-feel words can only be conveyed by providing the appropriate raw-feel experience." This is not true for complex raw-feel events, unless we assume the ineffability of simple ones. It is possible for me to explain to you the meaning of "visual image of a centaur" in language permitting you to recognize such a visual image (or even to create one); and this is true for the same reason that "centaur" in physical thing-language is verbally explicable without recourse to ostensive procedures. But I believe we can agree that it is true, in some sense, that certain "elementary" or "simple" raw-feel predicates cannot be thus verbally explained. I could probably convey the meaning of "orange" to a person whose visual history had been rigged so as to exclude any experiences in the (broadly specified) orange region, by telling him "Orange is a color between red and yellow, it's kind of a red-yellow mixture." But I could not do the same for yellow itself (which led some psychologists to reject a trichromatic theory of color on the ground that "yellow is a 'simple,' " although yellow can be matched on the color wheel by a suitable mixture of red and green). While there can be disagreement about many marginal cases, let us here grant for argument's sake that there exist *some* phenomenal predicates which designate raw-feel qualities of this "simple, unanalyzable, rock-bottom" kind. We would find it impossible to formulate in language a definition of the predicate in terms of more elementary concepts, concepts which taken separately are not synonymous (or even near-synonymous) with the predicate in question. A centaur is not simple—it has parts or components, it is "made out of" elements, "centaurhood" is a complex quality. Hence we can translate sentences about centaurs (and, derivatively, about centaur-images) into sentences about hooves, horses, men, etc. But if, say, *red* is a simple quality, we cannot thus reduce the predicate "red" to non-hue-characterizing expressions, enabling a learner to token appropriately in the future.

But even this is not quite true, taken literally as just stated. It would be possible to teach someone to token "red" appropriately, at least under many circumstances, by stating a special kind of semantic rule. We might, for instance, say " 'Red' designates the color of apples"; or " 'Red' designates the color you experience when light waves of wave length λ stimulate your eye." The availability of these verbal meaning-specifications is not philosophically important in the present context, because they will "work" only if the learner then proceeds to *put himself* into the described circumstances (i.e., to look at some apples or have a physicist stimulate his retina with wave length λ). So that our definition of "red" is, in effect, a prescription to him for learning obedience to the language-entry rule. Instead of training him ourselves—the usual way we enable children to acquire color words—we instruct him how to train himself. If we don't permit him to go through this intermediate process of self-tutelage, our definition of "red" will not enable him to token appropriately.

The ineffability of phenomenal simples can be trivialized by saying, "If a person has not been given the opportunity to learn a language, then, of course, he will be unable to speak it." As we discussed above in regard to the noninferential character of egocentric raw-feel statements, such statements are often *inferable* from other statements even though typically they are not, in fact, inferred. When tokened noninferentially, they have no "grounds" or "evidence" or "reasons" or "justification." The tokening occurs "on the basis of" a non-linguistic occurrence, i.e., the raw-feel event itself. The relation is (a) causal and (b) semantic, but the semantic rule is degenerate and uninformative. We learn to token in obedience to it by training, in the same way a dog learns to sit up or roll over. The required linkage is not intraverbal, as in rational inference or propositionally mediated knowledge;

the required linkage is between language and the nonlinguistic. And just as we have to be *trained* to perform intralinguistic transitions, so we have to be *trained* to perform language-entry transitions. From one point of view, since there is nothing philosophically earthshaking about the fact that a person cannot "give the right answer" to ($317 \times 48 = ?$) if he has never learned the multiplication table, why is it considered so very special that a person cannot "correctly name a raw-feel color quality" unless he has learned the color language?

IV. Is "Acquaintance" with a Raw-Feel Quality Cognitive?

This is perhaps an appropriate place to examine briefly what a person "knows" by virtue of having experienced a raw feel. Suppose that knowers K_1 and K_2 share knowledge of the Utopian scientific network, including, of course, the psychophysiology of vision and the psycholinguistics of color language. However, K_2 is congenitally blind and although he has recently undergone a corneal transplant, it is very shortly postsurgery and he has not as yet experienced any visual raw feels. Does K_1 know anything that K_2 does not know?

I have put this question to a number of "plain men" as well as to some not so plain. The natural tendency is to say that K_1 knows something that K_2 does not know, to wit, "*what red looks like*." Without taking vulgar speech or the theories it embodies as criteria of truth, I must confess that my own instincts in this matter are very like those of the plain man. It is pointed out by Professor Feigl that mere "living-through," "having," "experiencing" is not cognitive. While one might cavil at the persuasive definition of "knowing" presupposed by this exclusion, let us accept it for argument's sake. We may still question the conclusion that K_1 differs from K_2 only as to "*mere* experiencing." The raw-feel experiences of K_1 have been systematically correlated with color-word tokenings by himself and others during the course of his language-learning history, and he has been thoroughly trained to token "red" in accordance with the language-entry rule " 'Red' means *red*." He has not *merely* experienced the red and green hue qualities; rather he has coexperienced the red hue quality with "red"-tokenings and the green hue quality with "green"-tokenings. One reformulation of the plain man's notion that K_1 "knows how red looks, which K_2 does not" might be that K_1 *knows how to apply a language-entry rule which K_2 does not*. It is true that K_2 knows that there *is* a language-entry rule " 'Red' means *red."* He can token the rule itself. He can discuss the pragmatics of the rule (e.g., how people acquire the habit of obeying it, why he himself cannot as yet token in obedience to it, what the physical$_2$, functor values of the visual cortex are for brain-states Φ_r and Φ_g). He can describe precisely what needs to be done by him, or to him, so that he will be rendered capable of tokening in obedience to the rule in the future. Nevertheless he cannot token obedience to it at present. Is the ability to obey a semantic rule "cognitive"?

There are two related approaches to this question which seem to me to justify at least a tentatively affirmative answer. The first approach asks how it happens that a knower can know a semantic rule and yet be incapable of tokening in obedience to it. We can understand that people lie, make jokes, or mis-speak on a Freudian or other basis. But our knower K_2 *cannot* token color terms consistently with the semantic rules he "knows" (i.e., rules that he can token, and can fit into Utopian pragmatics). It is all very well to explain this oddity by causal analysis; it may even be somewhat illuminating to trivialize it as we did above by the statement "No one can speak a language he hasn't learned." But these psychological accounts, while perfectly correct as causal explanations, do not remove the oddity in its philosophic aspect. The more strictly philosophical question is: Granted that we

can easily explain K_2's status as a psychological consequence of his impoverished visual history, what, precisely, *is* that status with respect to the semantic rules "'Red' means *red*" and "'Green' means *green*"? The obvious answer, of course, is that ability to token a semantic rule, and the sincere motivation to emit other tokenings in obedience to it, does not suffice for doing the latter *unless one understands what the rule means*. Putting this affirmatively, we may say: "A person who understands the meaning of a rule can obey it if he wishes (provided, of course, that the action—in this case, tokening—is physically performable by him)." Therefore, if K_2 wishes to obey the rule but cannot, it must be that he does not understand it. I, like the plain man, am inclined to say that he does not fully understand it. To understand the semantic rule "'Rouge' means *red*" one must understand the meaning of "red." When blind K_2 tokens "'Red' means *red*" or "'Rouge' means *red*" he cannot fully intend all that K_1 intends by this utterance. He cannot, because "red" designates a hue quality and, as our plain man says, a blind person "does not know what red looks like." I am not able to push this argument any further, and I shall simply have to ask Professor Feigl whether he wants to say that a person who has never linked the words "red" and "green" to the hue qualities *red* and *green* can fully understand the semantic rules "'Red' means *red*" and "'Green' means *green*." Surely there is something intended by the seeing who enunciate this rule that is not intended by the blind who enunciate it?

It may be objected that our truism, "A person who understands the meaning of a rule can obey it if he wishes," was insufficiently qualified by the stipulation that the required action be physically performable by him. This way of putting it, focusing entirely upon the *action* (output) side, leaves open the question of his lacking sufficient information, and the latter is equally necessary for rule-obedience. If I am to obey the rule, "The king may not be put into check," it is insufficient that I am physically able to move the chessman; I must also know the position. Any rule which specifies an action appropriate to a condition will normally require for its consistent obedience that the action be physically performable *and* that the condition be knowable at the time of action-choice. And a semantic rule is such a rule—it makes specified tokenings licit conditional upon the occurrence of certain raw-feel events. The truism must be formulated more carefully so as to take this into proper account. A hard-nosed consistent identity theorist (tougher than Professor Feigl—say one like Professor P. K. Feyerabend) might argue thus: In order for K_2 to token in obedience to the rule "'Red' means *red*," he has to possess the information that he is experiencing a red-hued raw feel. We know (Utopian identity theory) that a red-hued raw feel is (*sic*!) a physical$_2$ state Φ_r in the visual cortex. Hence the semantic rule is "'Red' means Φ_r," and in order to be rule-obedient, one must token "red" if one's brain is in the Φ_r-state. Now the Φ_r-state is explicitly definable in terms of certain physical$_2$ functor inequalities, the satisfaction of which is ascertainable via the cerebroscope. Further, by means of the autocerebroscope (or an autocerebrophone, since K_2 has not yet learned visual form discrimination, so he couldn't read the television screen's symbols), K_2 can ascertain whether his brain is currently in state Φ_r or Φ_g, and he can do this upon the *first* postoperative occasion that red or green light enters his eyes. We have agreed that K_2 knows all the Utopian psychophysiology and the psycholinguistics that K_1 knows, so he will of course be able to understand the workings of the autocerebrophone and will token "red" (if, say, the instrument's high-pitched tone indicates his brain is in state Φ_r) and "green" (low-pitched tone indicating brain-state Φ_g) appropriately. It is therefore incorrect to say that K_2 "must not understand" the semantic rule, on the ground that he can't token in obedience to it. He can—provided he is given the necessary information.

Now this is an interesting move, and its use by an identity theorist testifies to his theoretical consistency. If the identity theory is taken literally—if a red-hued raw-feel event *is* literally, numerically identical with a physical$_2$ brain-event—then the autocerebrophone (or -scope, or -tact, or whatever) is a perfectly good epistemic avenue to that which is denoted variously in phenomenal, behavioristic, and neurophysiological languages. I rather suspect that this move is some kind of touchstone for an identity theorist's wholeheartedness; and I take note of the biographical fact that (in conversations) Professor Feigl finds the move somewhat distressing, as contrasted with Professor Feyerabend who sees it as *the* most straightforward approach and employs it aggressively and unapologetically.

Granted that K_2 could, by using the autocerebrophone, enable himself to obey the semantic rules even upon the first postoperative occasions of his experiencing red and green raw feels, what does this prove? Let us recapitulate the argument. We began with the principle that a person can, if he desires, obey a rule whose action-side is physically performable. We argued that just-operated knower K_2 cannot token in obedience to the semantic rule " 'Red' means *red*," even though he desires to, and even though he is able to perform the tokening act itself (i.e., he has long since learned how to *pronounce* "red" in the course of studying Utopian science). Therefore, we argued, he must not understand it. To this argument was offered the objection that the possibility of rule-obedience involves knowing the conditions which are to be compared to the rule's conditions, and that if K_2 were permitted to know the conditions (by using the autocerebrophone) then he *would* be able to token appropriately. Hence it has *not* been demonstrated that K_2 "does not understand the meaning of the rule."

This brings me to the second approach to "Does K_1 know something which K_2 doesn't?" The reason we—most of us, I mean—are not happy with the autocerebrophonic basis of K_2's rule-obedient tokening is that it is an alternative avenue, one which K_1 can dispense with because he has "more direct access." I do not expect this to disturb a good materialist like Professor Feyerabend, who of course views the situation as symmetrical in two indicator instruments: the tokening mechanism (biologically wired and learning-calibrated) and the autocerebrophone (scientifically built and calibrated), with the latter presumably being more reliable! Perhaps our dissatisfaction points the way to an argument not rebuttable by invoking the autocerebrophone?

Try this one: "If, given a shared system of propositions T and identical momentary physical inputs S_1 and S_2, a knower K_1 can predict a future external event E which K_2 cannot predict, then K_1 understands T more fully than K_2 understands it." Here T refers to the entire system of propositions known to Utopian science and shared by our two subjects, including the semantic rules. (It does not, however, include all particulars concerning the present experiment.) Inputs S_1 and S_2 are red light stimuli which we shine into the eyes of K_1 and K_2 (thereby giving the latter a red-hued raw feel for the first time). As we present this visual stimulus, we also give the following instructions to each subject: "I am going to shine this same light into the eyes of ten normal-visioned English-speaking subjects with the request to 'name the color.' What will they say?"

These instructions augment T somewhat, but equally for both subjects. The momentary stimulus inputs (while not identical) are "relevantly identical" in that they are clearly red (or can be made so by repeatedly stimulating K_2 before presenting the instructions). Yet we know that K_1 will be able to predict the verbal behavior of other subjects, whereas K_2 will not. Since both knowers have the same system T, the same instructions, the same

momentary input, but differ in their predictive ability, we conclude that K_1 understands something about T which K_2 does not. (I say "understands something," rather than "knows," because they can both token T's propositions, can both derive portions of T from other portions, can both engage in meta-talk about object-language parts of T, etc.) Of course the reason that poor K_2 cannot predict the verbal behavior of others is that merely being informed in our instructions that they will be stimulated by "this same light" does not entail that he will be informed what color the light is; hence he cannot move within the T-net-work to reach the semantic rule " 'Red' means *red*" or its pragmatic counterpart "English speakers token 'red' when they experience *red*." He knows these rules, but he—unlike K_1—cannot apply them to solving his cognitive problem. And there is, I submit, a very simple reason for his inability to apply them: *he doesn't fully understand them*. If he fully understood " 'Red' means *red*," he would know that the light entering his own eye was inducing red-qualitied raw feels in him; hence he would infer that "this same light" would have that effect upon others; and hence, finally, that they would token "red" in obedience to the rule.

Unless there is some flaw in this example and argument, it provides a refutation of the view that "knowing how to obey a language-entry rule is not cognitive." Because surely everyone will admit that he who can, by virtue of some extra "know-how," predict an external event which another lacking that "know-how" cannot predict has a genuine *cognitive edge* over the latter. Admittedly there is a considerable element of arbitrariness in where the line is drawn between "knowing that" and "knowing how," so that, whereas "knowing how" to solve physical problems by using calculus would be universally considered equally a "knowing that" and unquestionably cognitive, knowing how to obey linguistic rules is less clearly so. Our example suggests a fairly broad stipulation here, since we have seen that knowing *how* to obey a semantic rule can lead (via linguistic transitions) to a definite surplus of knowing *that*, i.e., rational prediction of a future external event. Knowing a language is conventionally considered cognitive; and we have good reason to follow conventional usage here. Not that anything philosophically critical hangs upon how tolerant a convention we adopt for labeling those marginal "knowings that" which are mainly "knowings how" as "cognitive." If a very narrow convention were adopted, it would still be true that possessing habits of semantic-rule-obedience gives one a cognitive edge, even if the habits themselves are classified as noncognitive. Suppose someone then maintains that raw-feel qualities can be deleted from the scientific world-picture, since that picture deals only with the cognitive, and the having of a raw feel is not cognitive. To this we reply the following:

1. A knower who has never experienced the raw-feel quality designated by a phenomenal predicate *Q* cannot mediate predictions (in the domain of descriptive pragmatics) which can be mediated by a knower who has experienced this raw-feel quality but who is in all other respects equivalently trained and informed.

2. This is because the inexperienced knower cannot employ the semantic rule " *'Q'* means *Q"* and the pragmatic rule "L-speakers token '*Q*' when experiencing *Q"* to make the necessary derivations.

3. His inability to employ these semantic and pragmatic rules in derivation chains arises not from ignorance of these rules, since he "knows" both of them.

4. Rather he is unable to employ them because he does not fully understand them.

5. Specifically, he does not fully understand " '*Q*' means *Q*" because he is unacquainted with the raw-feel quality *Q* which *"Q"* designates.

6. But if there is (exists, occurs) anything designated by "*Q*"—which must be so, if the inexperienced knower's predictive disadvantage is to be explained—then that something is left out of any world-account which does not include mention of *Q*.

7. A world-account which does not include *"Q"* when there is a cognitive edge favoring knowers who understand the meaning of "*Q*" is cognitively defective.

The basic point of all this, it seems to me, is really quite simple. What makes it seem difficult is our (understandable) phobia about "the given," "knowledge by acquaintance," "incorrigible protocols," and "the ineffable quale." If we shuck off these things (without adopting an untenable dispositionalism or logical behaviorism as our account of raw feels) the matter can be formulated as follows: Part of my knowledge consists of knowing when the people who speak "my" language use its various words. If I don't know this, if I cannot recognize a particular circumstance as being one for which the language has a certain expression, then there is clearly something about the language and its users which I do not know. But *what* is it about the language that I don't know? I know all of the language-transition rules; I even know—in a sense—the language-entry rules. But in another and critically important sense I do not fully "know" the language-entry rules, because there occur some situations legitimating a language entry in which others are able to make the prescribed entry but I am not. The common-sense explanation of this disability is that I cannot make the entry because I don't fully understand the meaning of some of the words involved. I am skeptical as to whether philosophese can improve fundamentally on this account of the matter.

Some have felt that this line of thought is adverse to the identity theory, but I am unable to see why. The identity theorist may agree that a knower who has experienced red-hued raw feels "knows" something, as shown by his ability to predict events under circumstances such that the prediction can only be mediated by the knower's habit of obedience to the language-entry rule " 'Red' means *red*." What bearing does this have upon the identity theory? We have adopted a tolerant conception of "the cognitive," and we therefore say that, while *mere experiencing* of a red-hued raw feel is not cognitive, such experiencing together with an appropriate tokening "red" (by another, or by oneself if thereupon reinforced by another) is a "cognitive" event in the derivative sense that it generates a "cognitive" disposition, to wit, the ability to token in conformity with the semantic rule; and he who knows how to obey the semantic rule "knows" something. Let it be further admitted that if K_1 can perform a rule-regulated language-entry under physical input conditions such that K_2 cannot, K_1 knows more than K_2 knows—in the present case K_1 knows "what quality the word 'red' designates." Does this prove, or have any tendency to prove, that "red" does not designate the physical$_2$ cortical state Φ_r? I am not yet raising the question whether (and, if so, how) a hue-quality term can designate a physical$_2$ cortical state (i.e., whether the "meaning" of "red" can *conceivably* be identified with any complex of physical$_2$ functor inequalities). I am only considering the anti-identity argument which arises from the K_1-K_2 cognitive edge. Can anything be inferred about ontological identity from the admission that a knower who is historically acquainted "by direct experience" with raw-feel quality *Q* can, by virtue of that acquaintance, recognize a new instance of it and token appropriately, whereas a knower who lacks such acquaintance "by direct experience" cannot do so?

How does identity theory handle this? In the reconstructed theoretical language of the identity theory, the semantic rule " 'Red' means *red*" can, of course, still occur; but there is

also a more informative semantic rule which provides an explicit definition of "red" in terms of physical$_2$ functor inequalities, i.e., the rule " 'Red' means the brain-state functor condition Φ_r." This semantic rule has been adopted (conventionally but nonarbitrarily) on the basis of a well-corroborated object-linguistic generalization that a red-hued raw-feel event *is* (in fact) a cortical state, characterizable in terms of functor inequalities represented by the explicitly defined symbol Φ_r. Within this revised language, we define *in usu*:

1. "Knower K experiences a raw feel y of red-hued quality at t" means that K's visual cortex is in physical$_2$ state Φ_r at t.

2. "Knower K is directly acquainted with the meaning of 'red' " means that he has experienced a raw feel of red-hued quality concurrently with a tokening of "red." (I neglect here the psychological refinements which would be necessary for completeness, e.g., we would have to distinguish between such "opportunities to become acquainted" and effectively "becoming acquainted.")

Then we have, in descriptive pragmatics, an empirical law which asserts, roughly,

3. A knower cannot token (extra-chance) in conformity with the semantic rule " 'Red' means Φ_r" unless either (a) He is directly acquainted with the meaning of "red," as in (2); in which case he tokens correctly by a language-entry only. Or (b) He is provided with other information as to the state of his brain or its physical$_2$ inputs; in which case he tokens correctly by a combination of *other* language-entries and language-transitions (i.e., he makes rational inferences).

We note in passing that each of these molar-level laws is in turn micro-derivable within the identity theory.

The appropriate tokening of "I see red" can occur so long as the knower's tokening mechanism is somehow brought under the causal control of his visual-cortical events. The causal chain from Φ_r to t_r, the "direct access" link, is available if he has learned the raw-feel language, but otherwise not. "Knowledge by acquaintance" consists, causally analyzed, in having established this intrabrain direct-access linkage.

Dualists often object to the identity theory on the ground that a person can know the meaning of "red" without knowing anything about brain-states. There is, at least to me, a troubling force to this objection if it goes to the intension of "(phenomenal) red," a point to which we shall return later. However, insofar as the objection lays its emphasis—as it commonly does—upon the pragmatic fact that a knower can consistently token in obedience to " 'Red' means *red*" without knowing anything about brain-states, I believe it has no force against the reconstructed identity theory. In order for this pragmatic fact to speak against the identity theory, we must show that the identity theory itself entails something to the contrary. But this is surely impossible to show in physical$_2$ terms. We would have to prove, within the identity frame, that in order for an organism's tokening mechanism to be differentially responsive (by tokening "red" versus "green") to the physical$_2$ states Φ_r versus Φ_g in the visual cortex, that same organism's tokening mechanism must also respond to the several physical$_2$ *component aspects* of Φ_r and Φ_g by tokening theoretical physical$_2$-language statements descriptive of these components. There is of course nothing about the laws of learning or the microphysiology of the brain which would even remotely suggest such a consequence. Insofar as the identity theory is, after all, a scientific theory—its substantive content being assertions in the physical$_2$ object-language—a refutation based upon the laws of tokening behavior must proceed within the theory's network and show therefrom that an entailed correlation among tokening dispositions is empirically false. The theory identifies

red-hued raw-feel events with Φ_r-states in the visual cortex. A Φ_r-state consists of a conjunction of component conditions, such as, say, "The second time-derivative of the proportion of simultaneously activated termini of the synaptic scales on cells of type M in cell-assemblies of form F lying within Brodmann's area 17 has a value $l_1 \leq d^2p/dt^2 \leq l_2$." Nothing within the theory need entail the consequence that a subject whose tokening system has been brought (through learned microconnections) under stochastic control of such a part-condition has also been trained to token so much as the word "derivative," let alone the whole collection of scientific statements involved in Φ_r. The tokening system is itself a complex entity, and its connections (wired-in plus conditioned) with the visual cortex are also complex. Tokening "red" is an event of immensely complex physical$_2$ composition, the components of which are not themselves tokenings. Nor are the microcomponents of the state Φ_r themselves raw feels. (Does anyone imagine, even if he is an identity theorist, that a single neuron "experiences the red quality" when that neuron undergoes a single discharge as part of its sustained firing contribution to the Φ_r-pattern?) There is just no reason to expect, on the identity theory's own grounds, that everyone who is able to token correctly in accordance with " 'Red' means *red*" should be able to token the theoretical statements contained in the explicit definition of state Φ_r. Given the causal laws and structural statements of the theory, we have no more reason to expect tokenings characterizing the microcomponent events or state-values than we have for expecting in the kinetic theory of heat that our thermometers should provide readings in terms of the mean free path or average velocity of the molecules in a bucket of hot soup. Therefore the fact of descriptive pragmatics that persons can "know what red is" (i.e., can know when they are themselves experiencing red-hued raw feels) without "knowing what a Φ-state is" (i.e., without knowing that their momentary brain-states fulfill certain functor conditions) provides no grounds for rejecting the identity theory *as being an inadequate causal account of the raw-feel tokening behavior*.

V. The Identity Theory and Leibniz's Principle

This causal analysis has its more strictly philosophical counterpart in the well-known qualification upon Leibniz's Principle, to wit, that it does not bind in intentional contexts. Actually our attempted refutation of the identity theory proceeds by ignoring this qualification, which in other philosophical disputes is routinely applied. The dualist is here arguing, in effect, that "K knows that he is experiencing a red raw feel," when conjoined with the identity theory's equivalence, "Experiencing a red-hued raw feel *is* having a Φ_r-state in one's visual cortex," entails "K knows that he is having a Φ_r-state in his visual cortex," a conclusion which is factually false (for all contemporary knowers, and for most knowers in Utopia). But the inference proceeds through substitution of the (red-hued raw feel = Φ_r-state) identity into the first statement "K knows..." a substitution via Leibniz's Principle in a forbidden intentional context.

Arguments against the identity theory which rely on Leibniz's Principle fall naturally into two classes, depending upon whether the allegedly unshared property is physical or mental. The commonest objection of the first class invokes one of the classic distinctions between the physical and the mental, to wit, that the former is "in space" and the latter is not. (Whether "in space" means space-*filling* or spatially *located* I shall not consider, although this refinement may be important in deciding upon a suitable convention for categorizing theories as monistic or dualistic.) Critics of the identity theory point out that

physical$_2$ brain-events are literally, anatomically, spatially *located in the head*; if Professor Feigl means what he says, and a red-hued raw feel is identical with a Φ_r-state, then since the Φ_r-state is in the head, he is committed to asserting that the red-hued raw feel is in the head. And the critic considers this a *reductio ad absurdum* on the ground that a raw feel is "obviously" *not* located in the head.

As stated above, I believe Professor Feigl is so committed by his theory, although I know him to feel a bit queasy about asserting "My phenomenal events are in my head." I shall attempt to show that this is an unobjectionable locution, and one which we *must* adopt if we take the identity theory seriously and literally.

Consider the phenomenal event which occurs in the ordinary negative afterimage experiment. Following upon fixation of a red circle in a booklet, I now fixate the gray wall twenty feet away and I "see there" a large, unsaturated, blue-green circle. While it is true that I do not *believe* it is "out there" (i.e., I do not assert that the causal ancestor of my current visual experience is a circular object on the wall), the "out there" character of the phenomenal event is strong, involuntary, and quite unmistakable.

Let us designate the color quality of this afterimage by the predicate "P," its apparent size by "*Q*," and its spatial ("out there") quality by "R." It does not matter whether "R" is 1-place or 2-place, since even if "my body" needs to be put in the second place, this "spatial" relation must of course be understood phenomenally, i.e., the red patch is *phenomenally external to my phenomenal body*. We use the 2-place predicate "L" as before, to mean "...is located (literally, spatially, anatomically) in the head of..."

Then when *a* is experiencing the afterimage we write, in our original notation for the psychophysical correlation-law and conjoining the identity statement:

$$(E!y)\,\Psi\,(a,y) \cdot P(y) \cdot Q(y) \cdot R(y) \cdot (E!z)\,L\,(a,z)\,\Phi\,(z) \cdot (y = z).$$

Which requires us to assert, substituting,

$$(E!y)L(a,y),$$

i.e., "The phenomenal event y is located in *a*'s head," which leaves us asserting the conjunction

$$(E!y)\,L\,(a,y) \cdot R(y),$$

i.e., "There is a phenomenal event located in *a*'s head and this event has the phenomenal property of seeming external to *a*'s head." This entailed conjunction is deemed contradictory by the critic, and is therefore taken to refute the identity conjunct (y = z).

But L(a,y) • R(y) is not self-contradictory, because the predicate L characterizes the physical$_2$ location of the phenomenal event, whereas the predicate R characterizes one of the phenomenal event's internal phenomenal properties. It is a phenomenal property of the afterimage to "appear out there," that is, the "externality" is *a quality of the experience*. In simpler notation, if *v* is a visual raw feel and *b* a brain-event, the identity theorist says

1. *v* is "out there (phenomenally)."
2. *b* is "in the head (anatomically)."
3. *v* and *b* are identical.

Spelling these out a bit more,

1′. The visual phenomenal event *v* possesses as one of its phenomenal properties an "out-there" quality.

2′. The brain-state *b* occurs in the head.

3′. The phenomenal event is the brain-state.

Hence we are required to say, "There occurs in the head a phenomenal event one of whose phenomenal properties is the 'out-there' quality." This kind of talk seems a bit odd, but what is there actually contradictory about it? R is a phenomenal predicate characterizing an experienced quality of the phenomenal event; whereas L is *not* a phenomenal predication and does *not* characterize the event as possessing or lacking experienced locus as a visual quality, but instead tells us where the phenomenal event is in physical$_2$ space. The critic relies on an allegedly clear truth, that

$$(x,y)[L(x,y) \supset \sim R(x,y)],$$

i.e., nothing can be "inside" and "outside" a person at the same time. But L means "physically inside" and R means "having the phenomenal quality 'outside,'" so nothing as to the predicability of R is inferable from the identity theory's universal predication of L. It goes without saying that the reconstructed theoretical language of Utopian identity theory will identify the phenomenal "out-there" property with certain physical$_2$ functor conditions in the brain-state, different from those identified with phenomenal hue, size, etc.

VI. Summary as to Nonsemantic Objections

Beginning with the autocerebroscopic thought-experiment, we have examined some of the commoner objections to the identity theory and, I believe, found them answerable. I have tried to show in what sense the theory can be genuinely empirical, by suggesting how Utopian neurophysiology might provide experimental results either corroborative or discorroborative of identity or dualism. In the light of the autocerebroscopic thought-experiment, I have briefly examined (with no claim to exhaustive treatment) several of the more familiar claims regarding raw-feel statements, especially those which are thought to militate against the identity theory. Most, if not all, of these claims would fail as refutations even if true, because they are claims about *knowledge* and therefore do not permit substitution via Leibniz's Principle. But I also tried to show that none of these claims is clearly correct, and that some are clearly incorrect. Raw-feel statements are, if I am right,

1. Noninferential normally, but not always, even when tokened egocentrically;
2. Corrigible;
3. Sometimes errant;
4. Indubitable, normally; but not always;
5. Private, contingently, but not nomologically or analytically;
6. Ineffable, when simple; but not in the sense of "incommunicable."

Finally I have discussed the question whether knowledge of raw feels by direct acquaintance is cognitive (I conclude that it is), and whether the occurrence of such knowledge without knowing the physical$_2$ nature of one's brain-events is ground for rejecting the identity thesis (I conclude that it is not).

The net result of these ruminations is, I think, to leave the identity theory in pretty good shape as an internally consistent hypothesis of "empirical metaphysics." The positive grounds for adopting it have not been discussed, and are too well known to require discussion. It must, I believe, be admitted that the evidence in its support, even within the limitations of current science and common sense, is fairly strong. Further, the purely *conceptual* difficulties we meet in trying to formulate even a sketch of ontological dualism (e.g., in getting clear "where Smith's non-physical$_2$ raw feels are" for purposes of their causally influencing events in Jones's brain) should discourage a scientifically oriented

theorizer from working along dualistic lines, at least until the more manageable monism runs into serious scientific difficulties.

VII. The Strongest Objection to Identity Theory: Semantic

I have left to the last the only objection (other than certain parapsychological phenomena) which I find continues to bother me somewhat. It involves the problem of intension—of "what sense-quality words *mean*"—in a way which is, I fear, rather too technical to be usefully discussed by a psychologist lacking the philosopher's union card. I shall therefore treat it only briefly in this place.

As a first approximation to formulating the difficulty, let the dualist advance the following: Your theory says that a red-hued raw feel is literally, numerically identical with a cortical Φ_r-state. According to this view, that which I experience as a red quality is—not "depends upon," "arises from," "correlates with," "corresponds to," "is produced by," but *is*—a complex of electromagnetic and electrostatic fields, alterations in potassium ion concentrations, disarrangements of membrane molecules, and the like. You have explained satisfactorily how it is that I can know when my brain is in this state, and how I can token "I am experiencing a red-hued raw feel" in obedience to the semantic rule " 'Red' means Φ_r" without knowing the physical$_2$ nature of the state Φ_r. You have resolved a paradox in pragmatics by combining a microcausal analysis of my learned tokening dispositions with a philosophical reminder that Leibniz's Principle does not bind in an intentional context. It is, as I now see, a mistake to attack you in the domain of pragmatics. But the problem has also a reflection in semantics. Surely you will concede that "red" and "green" designate raw-feel *qualities*. And these qualities do not appear anywhere in your theoretical reconstruction. When I examine the component expressions in the explicit definition of "Φ_r" I find physical$_2$ predicates and functors, but where are the red and green raw-feel qualities mentioned? They have got lost in the shuffle. I can't help feeling that, whatever you *have* done, you have *not* dealt with the problem we started to worry about, which was the nature of raw feels. Whatever else may be true of raw feels—and you have made it very doubtful whether anything else is uniquely true of them—at least it is true that they have qualities. And this, the sole distinguishing feature of raw feels, has been liquidated in the course of your analysis. I maintain that "red" designates a quality; that the *meaning* of "red" is this quality; and that there occur raw feels *of* this quality, i.e., "red-hued raw feel" denotes as well as designates. These qualities are in the world (if anything is!). But they find no place in your world-account. Therefore I must reject your account.

Professor Feigl himself has apparently some difficulties with the identity theory on this score. It might be said that his view is not entirely consistent, insofar as he takes the inverted-spectrum problem seriously. If the identity theory is correct, there can be no such *philosophical* problem, because Utopian science will presumably find out that your and my tokenings of "red" are correlated with the same kind of state Φ_r and that will settle the matter. If the identity theorist persists in philosophic doubt after that point in science is reached, or if he *adds* some sort of argument from "analogy," or appeals to "induction" or "determinism" or "parsimony" or "same cause, same effect," we know that he does *not* believe that a red-hued raw feel is literally identical with a Φ_r-state, since if it is, then there *is* no "other something" which requires analogy, etc., to infer.

Similarly, one may question Professor Feigl's emphasis upon the inter-subjective and causal features of scientific knowledge insofar as this knowledge is contrasted with our

"knowledge" of raw feels. If the identity theory is understood literally, raw feels are as much in the nomological network as anything else, and therefore they are part of the "public, inter-subjective" world-picture. As I have argued above, it is also necessary for them to be causally efficacious. If a red-hued raw feel *is* a Φ_r-state, and what distinguishes it from a green-hued raw feel is a difference in the physical$_2$ conditions, then the qualities *red* and *green* are represented by physical$_2$ expressions which are not nomologically different from those of nonpsychological science.

Within the frame of identity theory, can it be properly said that the occurrence of phenomenal qualities is "a contingent fact"? If we mean by this that the world could have been otherwise in its fundamental nomologicals, of course. Also if we mean that a different world of our world-family might have produced no living brains, and hence realized no Φ-states of raw-feel complexity. But what the identity theorist may *not* say is that the world could have been "just as it is, except for the absence of raw feels." It is not contingent (granting identity theory) that for every Φ_r-state there is a red-hued raw feel. This truth is analytic for an identity theorist. That the world could have contained no Φ_r-states is logically (and even nomologically) possible, but any world which does contain Φ_r-states contains red-hued raw feels. To return again to the kinetic theory analogy, a physicist who believes in kinetic theory can say:

1. Kinetic theory is empirical. It might be false; and, if true, it might have been false.
2. Even if kinetic theory is true, it might have been false that there ever existed such an entity as hot soup.

It is a contingent fact that the world has molecules in it. Given that there are molecules, it is a contingent fact that certain aggregates of molecules exist and have a certain mean kinetic energy. But what a physicist who believes in kinetic theory *cannot* say is:

3. It is a contingent fact that soup is hot when its molecules are moving rapidly.

The physicist who adopts kinetic theory has—until further notice—committed himself to a theory which demands obedience to the semantic rule " 'Hot' means *containing fast-moving molecules*." If subsequent evidence leads him to abandon the theory, he will again revise the semantic rules. But while he continues to hold the theory, statements like "The molecules of this hot soup are motionless" are analytically false. Similarly, for an identity theorist such statements as "Jones is experiencing no visual raw feel and his visual cortex is in state Φ_r*"* are analytically false. While disagreement exists among logicians as to which statements of a scientific theory are implicitly definitional of its constructs and which—if any—remain contingent even after adoption of the theory, I believe it is clear that what we may call *constitutive* statements always belong to the former class. Theoretical reductions in which complex physical$_2$ events are allegedly analyzed into their "parts" or "components," thereby permitting an explicit definition of the complex in terms of other theoretical primitives in the revised language, are constitutive in this sense. So that while one may adopt kinetic theory and (according to some) still view "Hot soup raises the mercury column of a thermometer placed in it" as contingent, as not frame-analytic, one may not adopt kinetic theory and yet view "Hot soup contains fast-moving molecules" as contingent. (I need hardly say that I am here using the phrase "hot soup" as a predicate of the physical thing-language and not as a predicate of a phenomenal sense quality.) The reduction of raw-feel statements to physical$_2$ statements is obviously constitutive within identity theory.

If the experienced properties of a raw feel were taken to *be* the raw feel instead of

properties of it, the identity thesis could be refuted easily by any knower who was directly acquainted with them. The refutation relies upon what Wilfrid Sellars calls the "difference in grain" between raw feels and physical$_2$ brain-states.[5] Suppose I am experiencing a circular red raw feel, large, clear, saturated, "focusing all my attention" upon its center. Granted that rapid fluctuations in attention can occur, that sensory satiation will take place, that the hue and "clearness" at the circle's edge may differ from those at its center; nevertheless, the most careful and sophisticated introspection will fail to refute the following statement: "There is a finite subregion ΔR of the raw-feel red patch Ψ_r, and a finite time interval Δt, such that during Δt *no property* of ΔR changes."

The properties of a phenomenal event being the properties it "appears to have," we state the above in phenomenal object-language and avoid any unnecessary reference to knowing. This is important because the desired refutation utilizes Leibniz's Principle. We say of the physical$_2$ state Φ_r (by substituting Φ_r for Ψ_r as allowed by the theory): "There is a finite region ΔR of the brain-state Φ_r and a finite time interval Δt, such that during Δt no property of ΔR changes." But this, as even pre-Utopian neurophysiology shows us, is factually false. The region ΔR and the interval Δt are not infinitesimals, they are finite values taken small enough to satisfy the raw-feel statement of constancy. Thus, during, say, 500 milliseconds, the 5° region at the center of my phenomenal circle does not change in any property, whereas no region of the physical brain-event can be taken small enough such that *none* of its properties changes during a 500-millisecond period.

This "grain" objection cannot, of course, be answered by saying that a property of the brain-state, such as an average value of a certain complex physical$_2$ functor, remains constant during Δt, analogizing to the relation between macrotemperature and molecular motion (compare Feyerabend, 1963, page 53). The phenomenal assertion is stronger than this rebuttal can meet, because it says "…no property changes…" not "…some property remains constant…"

This "grain" argument seems to me to provide a clear refutation of the identity theory, provided we identify "the raw-feel event" or "the phenomenal entity" with the experienced circular red patch. But we need not do this, and the identity theorist will of course not do it.

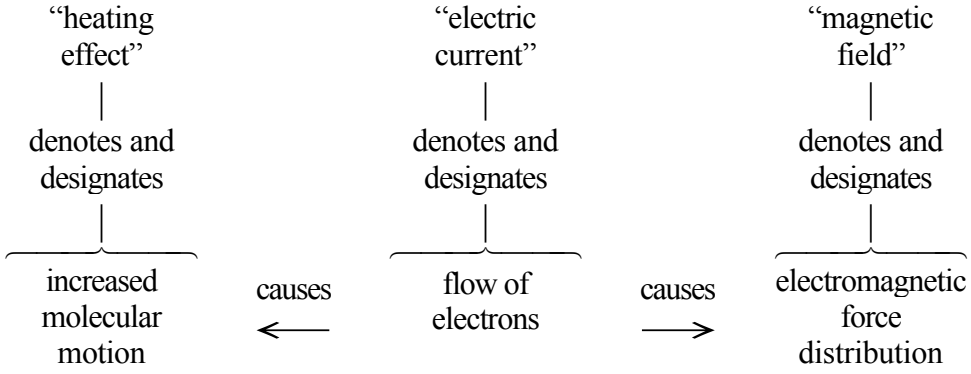
---

[5] I am not sure that I correctly understand Sellars's use of the term "grain," but I learned the term, and this objection to the identity theory, in discussions with him. Roughly put, "grain" refers to an admittedly vague cluster of properties involving continuity, qualitative homogeneity, unity or lack of discrete parts, spatiotemporal smoothness or flow, and the like, which many raw feels possess in ways that their corresponding physical$_2$ brain-states do not. Thus a small phenomenal red patch is typically experienced as a continuous, homogeneous expanse of red hue. The identity theory makes the phrase "phenomenal red patch," in the revised theoretical language, refer both to this entity and to that "gappy," heterogeneous, discontinuous conglomerate of spatially discrete events that are described in a physical$_2$ account of the brain-state Φ_r. The issue raised is similar to Eddington's problem about "which table is the real table"—the solid object of ordinary experience or the inferred entity of physical theory, mostly empty space sparsely occupied by gyrating electrons? Whereas Eddington's problen is fairly easily dissolved by proper linguistic analysis (both tables are real, being the same table, and the macrodispositions being causally analyzed in terms of the microstructure), its analogue in the identity theory is more refractory. The objection is termed "semantic" because it is, fundamentally, based upon an alleged radical difference in (*intensional*) *meanings* (= designata) of phenomenal and physical$_2$ predicates, taken together with Leibniz's Principle.

He will instead speak of the entity as *having the properties* red, circular, saturated, etc. He thinks of the denotatum as a *tertium quid*, whose existence becomes "known" to science either via the internal linkage to a tokening mechanism ("knowledge by acquaintance") or via the external linkage to the cerebroscope (hetero- or auto-, it doesn't matter). From Professor Feigl's standpoint, the identity of a raw-feel event as known "from the inside" with a raw-feel event as known "from the outside" is of the same sort as the identity between the morning star and the evening star. He relies upon the principle that identity of denotata does not imply identity of designata.

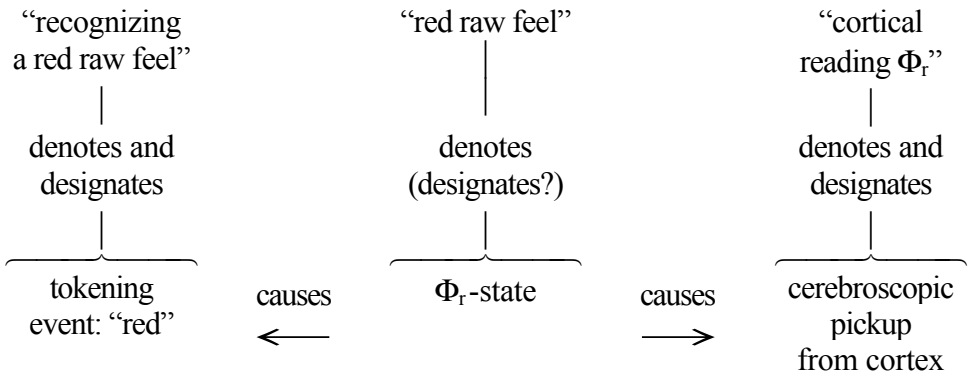
Many find this analogy unsatisfying, myself among them. Let me try to say why it bothers me. Since the morning star example involves an individual, it will be preferable to use another of Professor Feigl's own examples, to wit, the diverse indicators of an electric current. Passage of an electric current is associated with several indicators of its occurrence, two among them being a heating effect and an electromagnetic field effect. A temperature rise in the conductor or its vicinity is an "avenue of knowledge" to the *tertium quid*: electric current, as is the deflection of a compass needle near the wire. Professor Feigl wishes to say that the phenomenal quality *red* (known "by acquaintance," "directly," "from inside") and the cerebroscopically indicated neurophysiological facts (known "by description," "indirectly," "from outside") are related to the one brain-event in the same way that thermoelectric and electromagnetic indicators are related to the one electric current.

In critically examining this analogy, I first take note of the fact that a temperature rise and a compass deflection are events (or states) distinct from the denotatum of "electric current." The entity *designated and denoted* by "electric current" is a movement of electrons through the conductor. This entity is nomologically linked to two other sorts of events, but theory does not identify them with *it* any more than with each other. We say that *when* and *because* the electrons move through the conductor, the latter's molecules increase their average velocity (which we detect by means of a thermometer) and an electromagnetic field surrounds the conductor (which we detect by means of a compass). There are *three* designata, and *three* denotata, thus:



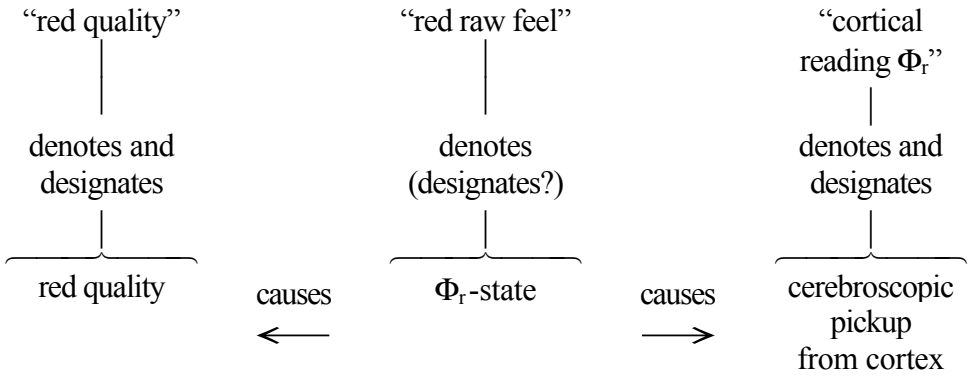
I note also that we do not consider our theory complete unless it provides a theoretical (derivable) answer to the question why the passage of electrons can be detected by thermometers and compass needles, i.e., why the indicators work.

Now if Professor Feigl took the view that the only "inside indicator" was the tokening event itself, his analogy would be strictly accurate. We would have the following:



("Red raw feel" only *denotes* the Φ_r-state as tokened egocentrically, although it also *designates* the Φ_r-state as tokened by a sophisticated Utopian believer in the identity theory.) The above diagram represents Professor Feyerabend's view of the identity theory, since he rejects the claim that phenomenal quality words designate anything other than the vaguely understood Φ-states to which they are linked by language-entry habits.

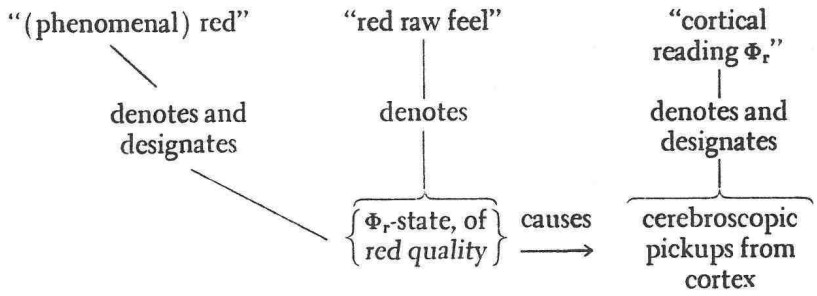
However, the above diagram does not represent Professor Feigl's view of the matter, since he admits—nay, insists—that phenomenal predicates *do* designate something, and that their "meaning" is known to us by acquaintance. That is, the *red quality* is itself an indicator, the tokening "red" being still farther along in the indicator chain. If a person fails to token "red" on certain occasions when he experiences a red raw feel, there nevertheless occurs an instance of the red quality, according to Professor Feigl's view. How is this interpretation to be diagramed? Following the analogy, we have:



But surely this won't do. We have a something called "red quality" being *produced by* the Φ_r-state. Such an interpretation is objectionable on at least two counts. First, qualities are *of* existents, they do not exist on their own. Secondly, if there is a red quality which is "produced by" the Φ_r-state, the physical$_2$ reduction sought by the identity theory is endangered. "Red" as a phenomenal predicate is not to be found among the physical$_2$ functors and predicates of physical theory. If there is a property designated and denoted by the quality-word "red" which does not "belong to" (or is not "part of") the physical$_2$ Φ_r-state, then the identity theory is not—literally—an *identity* theory; it is only a weaker claim that all mental events are physical$_1$, i.e., nomologically linked to physical$_2$ occurrents and dis-

positions. This claim is, of course, quite compatible with dualism (interactionistic or epiphenomenalistic, assuming that the latter can be made consistent).

But if there is no numerically distinct *entity* (even, so to say, a "short-lived continuant") to which the Φ_r-state gives rise, then the quality-word "red" must denote *a property of the* Φ_r*-state itself*. The causal and semantic situation would then be represented thus:



In this interpretation, "red-hued" is a property of the physical$_2$ brain-state Φ_r. But since, as Professor Feigl holds, "red" as a phenomenal predicate is absent from the list of fundamental predicates and functors of theoretical physics—speaking object-linguistically, among the elementary properties and dimensions of the world there is no such thing as a red quality—how is it theoretically possible for *red* to be a property of the physical$_2$ brain-state Φ_r? So far as I can see, the only way this would be possible is if "red-hued" designates a *configural property* of the complex physical$_2$ state Φ_r. That is, "red-hued" designates a complex conjunction of physical$_2$ functor inequalities, analogously to the way in which "hot" (physical thing-language, not phenomenal language!) designates a summary statement about molecular motion, or "charged" designates a configural statement about the distribution of elementary particles on the plates of a condenser.

Such an analysis, forced upon us by our determination to maintain a genuine *identity* thesis, brings us back again to the counterintuitive difference in "grain" between the intension ("by acquaintance") of "red hue" and the intension ("by description") of "Φ_r" formulated in terms of physical$_2$ functors. But in the course of returning full circle to that objection on the basis of meanings, we have, I fear, precluded Professor Feigl from one kind of defense. Initially, he was able to answer the "grain" objection by distinguishing *property* from *entity* and *designatum* from *denotatum*. This rebuttal is no longer available to him because now the "grain" objection is reiterated at the level of *properties* rather than at the level of occurrents or continuants themselves. Can anyone who knows the red quality "by acquaintance" really allow, on the basis of any theoretical reconstruction, that this quality *is* a configural property of physical$_2$ components?

I must confess that I do not know how to put this question in a less "intuitive" form, which leaves me in an unsatisfactory philosophical position vis-à-vis a radical materialist like Professor Feyerabend (who, when I put it to him, looks me right in the eye and answers "Yes"). I am impelled to pursue the argument in terms of "understanding the equivalence," but I do not know how to do this with any philosophic rigor. It strikes me as very odd that I could fully understand the intension of the phenomenal predicate "red-hued," this predicate designating a configural physical$_2$ property, and could also fully understand the intension of the physical$_2$ predicate "Φ_r"; and yet not "understand why" they designate the same *property*. This seems to me radically different from the situation

where I can understand two designating expressions without knowing that only one denotatum satisfies them. In the morning star-evening star example, it is easy to see why I can understand the meaning of each expression and yet be uninformed as to the factual identity. We deal there with the difference between propositional functions and the (otherwise unknown) individuals whose names or definite descriptions can be filled into the variable positions; whereas in the present case we are dealing not with values of the variables but with the intension of the predicates. Not to understand that "author of *Waverley*" denotes the same entity as "author of *Rob Roy"* is compatible with understanding both expressions, because understanding the identity involves knowledge of *facts*. Similarly, understanding the meanings of "specific etiology of paresis" and "specific etiology of tabes" is compatible with ignorance that their denotata are identical, again because factual information is involved. But the identity theory—if held consistently—identifies the *property* designated by "red" with the complex physical$_2$ property designated by "Φ_r."[6] I do not see how one can fully understand the intension of two property expressions which designate the same property—one expression being shorthand for the other, which is complex—without understanding their equivalence.

And, be it noted, this lack of understanding can be expected to continue even *after* a claim of identity has been made within the revised theoretical language. Consider a Utopian neurophysiologist who is not blind and who has learned to speak the ordinary color language, but who has *not studied either psycholinguistics or psychophysics* (i.e., he knows all about the fundamental nomologicals—the laws of physics—but he does not know those structure-dependent quasi-nomologicals that are provided by the contingent facts of his verbal culture and the characteristics of organic transducers). He does not, therefore, know *all* of Utopian science, because that corpus includes such disciplines as descriptive pragmatics and the psychophysiology of vision. He does, however, know cortical anatomy and neurophysiology. We provide him with the following information about experimental particulars:

1. *Stimulus*: A bright, saturated red light is made to fall upon his retina.
2. *Proposition*: "A light stimulus, possibly but not necessarily of this kind, was applied to Subject X, a physically normal individual."
3. *Proposition*: "The cerebroscopic reading off X's visual cortex under that stimulation was Φ_r."

Query: "Was X stimulated with the same light you were?"

What will be our Utopian's epistemic situation given these inputs? Having learned the habit of obedience to the language-entry rule (i.e., knowing the meaning of "red" by acquaintance) he can correctly token "I experience *red*" and, via nomologically legitimated language-transitions, can thereby also token correctly, "If X was stimulated by this same light, X experienced *red."* He also knows that X's visual cortex was in state Φ_r as a result of whatever light stimulation he received. There is nothing about this physical$_2$ description of X's brain-state that he does not "fully understand." Why then can he not tell us whether the state "Φ_r" matches "experienced *red"?*

The situation here is different from that of our congenitally blind Utopian, who could

[6] Professor Feigl does not, of course, assert this identification of properties; quite to the contrary, he wishes to deny it. But I hope that my discussion up to this point has shown that a genuine, consistent identity theory cannot escape such an identification of the property *phenomenal red* with the property Φ_r.

not solve his cognitive problem because he did not *know the language*. It is also different from that of a non-Utopian, who does not know all about the fundamental nomologicals (or, at least, does not fully understand the microphysiology represented by the physical$_2$ description Φ_r). Our present subject "knows all there is to know" about the state Φ_r which occurred in the brain of X. He "knows all there is to know" about the phenomenal quality *red,* including how to designate it (by "red"). If "red" is a shorthand expression for the physical$_2$ configural property Φ_r, he should be able to say that the states are the same. But, of course, he cannot.

Nor will it do for the identity theorist to complain that we have unfairly rigged the thought-experiment by withholding crucial information, when we disallowed him knowledge of psycholinguistics and the visual-system transducers. It is not relevant to the *nature* of state Φ_r that it is induced by the "usual" (retina $\rightarrow$ … $\rightarrow$ lateral geniculate $\rightarrow$ ...) causal chain. Nor is it relevant to the rule " 'Red' means *red*" that people historically learn it in a certain way. Our subject knows the rule, and he knows that X knows the rule. He knows what "red" *means*; and he *knows that it means the same to* X *as to himself.* And he knows the configural state of X's brain. If *red* is, literally, nothing but a configural property of the Φ_r-state, it is very strange indeed that he cannot match "red" with "Φ_r" so as to postdict X's stimulus input.

And what about after we inform him? Even *after* he knows that phenomenal red consists of (*sic*!) the physical$_2$ configuration Φ_r, will this seem in any way "appropriate" or "comprehensible" to him? It is always dangerous to anticipate science, especially in the negative. But I cannot conceive that any theoretical reconstruction involving the fundamental physical$_2$ functors would enable me to "see how" the red phenomenal quality *consisted* of such and such a configuration of fields, ions, disrupted neuron membranes, and the like. And I do not believe that Professor Feigl envisages any such possibility either.

I have not been able to formulate the psychophysical correlation-law and, as a separate claim, the identity thesis, by means of a notation essentially different from that of Section I; nor has Professor Feigl suggested any such (personal communication). It would seem not only natural but unavoidable, in meeting Professor Feigl's own conditions on the epistemological status of the identity thesis, that our notation should represent the entities to be identified—the raw-feel event and the brain-state—by bound *variables*; whereas the raw-feel qualities predicated of the former, and the physical$_2$ functor conditions characterizing the latter, should occur in the role of descriptive constants. I think that the notation brings out more clearly than words what is intuitively unsatisfying about the morning-star kind of analogy. It also shows why Professor Feigl's invocation of the designatum-denotatum distinction, while important in forestalling certain alleged refutations relying upon unnecessary puzzles in pragmatics, does not quite succeed in clearing matters up. Consider a critic who attempts to refute the identity thesis by saying, "A psychophysical correlation-law presupposes that there are *two* '*somethings*' being correlated; you admit that there is such a law, and that it is empirical. How then can you turn about and assert that there is only *one* '*something*' without contradicting these very statements which constitute your main scientific ground for adopting the identity thesis in the first place?" Contemplation of the logical form of our correlation-law provides the answer to this objection. We have

$$(x,y)\ \Psi\ (x,y) \cdot P\ (y) \rightarrow (E!z)\ L\ (z,x) \cdot \Phi\ (z)$$
$$(x,z)\ L\ (z,x) \cdot \Phi\ (z) \rightarrow (E!y)\ \Psi\ (x,y) \cdot P(y)$$

which shows that a raw-feel event y is characterized by phenomenal hue quality P and related (Ψ) to a person x; a brain-state z is characterized by the complex physical$_2$ functor condition Φ and related (L) to the body of that same person. The correlation-law asserts that the necessary and sufficient condition for a person to be the experiential locus of a raw-feel event having property P is that his brain should be in state Φ. Obviously there is nothing about this formally which estops us from a subsequent assertion that the event and the state are identical. The critic cannot complain of a "shift" here from speaking about two entities to speaking about only one, unless he is prepared to maintain that there is something impermissible logically about saying "One individual wrote *Waverley*," and "One individual wrote *Rob Roy*," following these with "And these ('two') individuals are the same ('one')." The psychophysical correlation-law merely informs us that whenever there is an entity satisfying one pair of prepositional functions (P,Ψ), there is a unique entity which satisfies another pair of propositional functions (L,Φ); it leaves open the question as to whether one or two numerically distinct entities are involved. So Professor Feigl is correct in answering *this* criticism in terms of the usual distinction between designatum and denotatum.

But the notation also shows clearly why such an attack is misconceived, being aimed at the *variables* instead of at the descriptive constants. If Sellars is right in seeing an insuperable objection to the identity thesis *even on present knowledge* to be the difference in "grain" between phenomenal events and physical$_2$ brain-states, then the proper focus for attack is not the conjoined identity claim "(y = z)" itself, but (via Leibniz's Principle) its consequence, to wit, that

$$(f)\,(fy \equiv fz)$$

from which we conclude that, given the correlation-laws,

$$(y)(Py \supset Pz)$$

i.e., the physical$_2$ brain-state must possess the red hue quality. And since the predicate "P" designating the red hue quality is not to be found among the component physical$_2$ functor conditions ϕ_1, ϕ_2, … ϕ_m which jointly constitute the explicit definition in our theoretical language of the brain-state property Φ, we must conclude that "P" designates the configural physical$_2$ property Φ itself. Quite apart from the possibility that property P is causally inefficacious (epiphenomenalism), or that "P" does not designate the same quality in Professor Feigl's language and in mine, we have his insistence that "P" does designate something that exists, that there *are* phenomenal qualities "in the world." Nor does it help him to argue—if indeed he can successfully—that they are not includable within the intersubjective world-network of physical$_1$ science. Be that as it may, he is concerned about the mind-body problem qua philosopher (if not "of science," then "as empirical metaphysician") because he holds that phenomenal predicates have a referent, that they denote something, namely, the raw-feel qualities themselves. He stoutly maintains that he has raw feels and is acquainted with their qualities; he cheerfully admits that others have them too. So we are initially agreed that the phenomenal predicate "P" does refer to an existent quality. Granted that there is such a quality, the identity thesis entails via Leibniz's Principle that it be literally attributable to the brain-state. Hence in the reconstructed theoretical language of identity theory, we should adopt the semantic rule

"P" means Φ

which I, like Sellars, find it quite impossible to make myself genuinely intend.

Can the "grain" objection be restated in terms of properties? Yes, it can, and in a tight, simple form which is unavoidable except by denying its premises. If "Sim(...)" is a second-type one-place predicate designating the property *simple* (a property of first-type properties) we assert:

$$\begin{aligned} &\mathrm{Sim}(P) \\ &\sim \mathrm{Sim}(\Phi) \\ &\therefore P \neq \Phi \end{aligned}$$

by Leibniz's Principle applied to properties. The only trouble with this direct hammer-blow is that "Sim(P)" itself is not provable, although it has a strong intuitive obviousness to most (but, alas, not all) thinkers. And while its intuitive claim upon me (and, interestingly, upon Professor Feigl) is compelling, I am hardly prepared to insist that English or epistemologese has a clear language-entry rule about *simplicity* which is violated by a denier of the premise "Sim(P)."

Another approach to rigorous formulation of the "grain" argument, also involving predicates of higher type, relies upon the alleged nontransitivity of the "equal," "nondiffering," "indiscriminable," or "indistinguishable" relation among phenomenal qualities. Some have argued that "indistinguishability" is transitive for physical$_2$ properties but is nontransitive for phenomenal properties. The highly technical issues involved in that allegation are beyond the scope of this paper, and I shall content myself with making two critical observations. First, the commonly assumed nontransitivity of phenomenal "equality" rests upon experimental facts interpreted via an arbitrary definition of the difference threshold (e.g., the old 75 per cent criterion, which is wholly without logical, psychometric, physical, or physiological justification). Secondly, the autocerebroscopic thought-experiment and its attendant theoretical speculations should have made it clear to the reader that it is *not* absurd, meaningless, or self-contradictory to say that two raw feels "seem equal but are not," the reason being that "seeming," when carefully analyzed, invariably turns out to be a matter of tokenings *or other intervening or output events*. I must emphasize that no dispositional or logical-behaviorist analysis of mind is presupposed in saying this; it must be obvious that I reject all such analyses. Nor is the identity thesis or any variant of "materialism" presupposed. I am simply insisting that "seeming equal" is a state, process, event (whether physical$_2$ or not) which, while a causal descendant of the raw-feel event and correlated with the latter's properties, is certainly not to be identified with the raw-feel event or its properties. "These two phenomenal [*sic*!] greens seem equal to me," which expresses a *judgment about* experience that is numerically distinct from the experience, does *not* entail phenomenal equality. Taking these two considerations jointly into account, I do not believe we are compelled by the available psychological evidence to assert that "equality" is nontransitive for phenomenal properties.

I warned the reader at the start of this concluding section that I was not competent to present a rigorous philosophical objection on purely semantic grounds, and I am acutely conscious of not having done so. It remains my conviction that there is something fishy about the identity thesis when interpreted literally, i.e., as a genuine *identity* thesis. Professor Feigl must, I think, make his mind up as to whether or not there are any raw-feel qualities, i.e., whether phenomenal predicates denote anything. I do not believe that he has solved the basic problem by emphasizing the designatum-denotatum distinction, because this only takes care of the relation between the tokening of phenomenal predicates and the neurophysiological readings off the visual cortex. He—unlike Professor Feyerabend—

maintains that the reference of phenomenal predicates is to raw-feel *qualities*. This insistence creates a dilemma for him: if these qualities are other than complex physical$_2$ functor conditions, then the "identity thesis" is misleadingly named; for there is something in the world—and a causally efficacious something at that—which is not reducible to the theoretical entities of physics. Alternatively, if these qualities are *not* other than physical$_2$ functor conditions, they must be configural combinations of the latter. It does not seem to me that they can possibly *be* that, but I leave it to a competent philosopher to prove what to me is only intuitively obvious as a matter of "grain."

If the identity of raw-feel red with a physical$_2$ configural property can be shown impossible upon rigorous *semantic* grounds, the identity theory is demolished; and the outcome of an autocerebroscopic experiment is thereby rendered partly irrelevant and partly predictable. Per contra, if the equivalence of "phenomenal red" and "brain-state Φ_r" is free of semantic difficulty, then I think it must be admitted that the identity theory is in a very strong and easily defensible position. In particular, I have tried to show in this paper that some of the commonly advanced nonsemantic refutations of it are without merit.

REFERENCES

Bohr, Niels. (1934) *Atomic theory and the description of nature*. New York: Macmillan.

Carnap, Rudolf. (1950). Empiricism, semantics, and ontology, *Revue Internationale de Philosophie*, 11:20-40.

Carnap, Rudolf. (1952). Meaning postulates. *Philosophical Studies*, 3:65-73.

Eccles, J. C. (1951). Hypotheses relating to the brain-mind problem. *Nature*, 168:53-57.

Eccles, J. C. (1953). *The neurophysiological basis of mind*. Oxford: Oxford University Press,.

Eddington, A. S. (1939). *The philosophy of physical science*. Cambridge: Cambridge University Press.

Feigl, Herbert. (1958). The 'mental' and the 'physical.' In Herbert Feigl, Michael Scriven, & Grover Maxwell (eds.), *Minnesota Studies in the Philosophy of Science, Vol. II* (pp. 370-497). Minneapolis: University of Minnesota Press.

Feyerabend, P. K. (1962). Explanation, reduction, and empiricism. In Herbert Feigl & Grover Maxwell (eds.), *Minnesota Studies in the Philosophy of Science, Vol. III* (pp. 28-97). Minneapolis: University of Minnesota Press.

Feyerabend, P. K. (1963). Materialism and the mind-body problem. *Review of Metaphysics*, 17:49-66.

Howells, T. H. (1944). The experimental development of color-tone synesthesia. *Journal of Experimental Psychology*, 34:87-103.

Jordan, Pascual. (1955). *Science and the course of history.* New Haven, CT: Yale University Press.

Lachs, J. (1963). The impotent mind. *Review of Metaphysics,* 17:187-199.

London, I. D. (1952). Quantum biology and psychology. *Journal of General Psychology*, 46:123-149.

Maxwell, Grover. (1961). Meaning postulates in scientific theories. In Herbert Feigl & Grover Maxwell (eds.), *Current Issues in the Philosophy of Science* (pp. 169-183). New York: Holt, Rinehart, and Winston.

Meehl, P. E. (1950). A most peculiar paradox. *Philosophical Studies,* 1:47-48.

Meehl, P. E., Klann, H. R., Schmieding, A. F., Breimeier, K. H., & Sloman, S. S. (1958). *What, then, is man?* St. Louis, MO: Concordia Publishing House.

Meehl, P. E., & Sellars, Wilfrid. (1956). The concept of emergence. In Herbert Feigl & Michael Scriven (eds.), *Minnesota Studies in the Philosophy of Science, Vol. I* (pp. 239-252). Minneapolis: University of Minnesota Press.

Popper, Karl R. (1962). *Conjectures and refutations*. New York: Basic Books.

Popper, Karl R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Reichenbach, Hans. (1938). *Experience and prediction*. Chicago: University of Chicago Press.

Scriven, Michael. (1953). The mechanical concept of mind. *Mind*, 62:230-240.

Sellars, Wilfrid. (1948). Concepts as involving laws and inconceivable without them. *Philosophy of Science*, 15:287-315.

Sellars, Wilfrid. (1953). Is there a synthetic a priori? *Philosophy of Science*, 20:121-138.

Sellars, Wilfrid. (1947). Pure pragmatics and epistemology. *Philosophy of Science*, 14:181-202.

Sellars, Wilfrid. (1948). Realism and the new way of words. *Philosophy and Phenomenological Research*, 8:601-634. Reprinted in Herbert Feigl & Wilfrid Sellars (eds.), *Readings in Philosophical Analysis* (pp. 424-456). New York: Appleton-Century-Crofts, 1949.

Sellars, Wilfrid. (1953). A semantical solution of the mind-body problem. *Methodos*, 5:45-84.

Sellars, Wilfrid. (1954). Some reflections on language games. *Philosophy of Science*, 21:204-228.

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, 52:270-277.

pdf by LJY
January 2015