The Imperative of Responsibility

In Search of an Ethics for the Technological Age

Hans Jonas

Translated by Hans Jonas with the Collaboration of David Herr

ISBN 0-226-40596-6

The University of Chicago Press
Chicago & London
1984

"Interaction" 222,213,226,243n11 — the crux of the matter.

Appendix to Chapter 3 (III.5)
Impotence or Power of Subjectivity
A Reappraisal of the Psychophysical Problem

Historical Prologue

I begin with an episode from the history of science. In 1845, three young, enthusiastic physiologists, pupils of the great Johannes Müller in Berlin, made a formal pact to fight the view, hitherto regnant and called "vitalism," that life involves forces other than those found in the interaction of inorganic bodies. They were Emil du Bois-Reymond, Ernst Brücke, and Hermann von Helmholtz—all to become very famous. Du Bois-Reymond and Brücke between them even pledged with a solemn oath that they would establish and compel acceptance of this truth: "No other forces than the common physical chemical ones are active within the organism."[1]

All three kept their vow throughout their lives, with spectacular scientific success, which in turn helped to make their proposition an article of faith, and "vitalism" a discredited cause, throughout the life sciences. What escaped them was the fact that by making *a* pledge, no matter which, they already contradicted or acted counter to the particular content of *this* pledge. For they did not bind themselves to what is not a matter for decision at all—namely, to let the molecules of their brains now and hereafter take their causally prescribed courses and allow them to determine their thought and speech (if they do that anyway); but they bound themselves to remain faithful in the future to a present insight, thus by implication declaring at least *their* subjectivity to be master over their conduct. In the mere fact of a vow, they credited something totally *non*-physical, their relation to truth, with just that *power* over their overt behavior which the *content* of their vow on principle denied. Making a promise, with faith in the ability to keep it and the equally implied alternative of *not* keeping it, does admit into the chart of reality—if not into

scientific discourse—a force "other than those found in the interaction of inorganic bodies." "Fidelity" would be such a force, as every other rule of mind over behavior, that is, body. And yet they were right to exclude the psyche from scientific discourse in their program for physiology. This is the genuine antinomy of the situation, which no one can evade.

As to the deuced psyche itself, or consciousness, or subjectivity, which has the gall to be there and is even less deniable than the being-there of bodies in space, and what to make of its copresence and interrelation with the physical basis, one of the three signers of the pact, du Bois-Reymond, coined twenty-seven years later in a famed lecture "On the Limits of the Knowledge of Nature" the much-maligned phrase that we not only do not know (*ignoramus*), but on principle never shall know (*ignorabimus*).[2] The indignation of many scientists at this "defeatism" was great, but I suspect that he was right *in the terms of natural science*. And philosophy, since Descartes, cannot boast of much better success. Still, this skeleton in the closet of modern science will insist on rattling and disturbing our sleep. Reason, as Kant perceived, cannot leave it alone—nor can it, contrary to Kant, acquiesce in a sheer antinomy. Even with the *ignorabimus* waiting implacably at the end, we might gain some ground from *complete* ignorance by at least clearing the issue of distorting and obstructing misconceptions—and in the process perhaps even achieve something better than the jarring clash of incompatibles, which after all is itself a positive assertion and as such can be in error.

I shall first try to articulate the hoary "psychophysical problem," then discuss the "epiphenomenon" formula with which it has been bent one-sidedly to the purported demands of natural science or materialism, and last I shall try my hand at a conjecture which better fits the bifurcated data of our experience, neither of which ought to be disenfranchised in deference to the other.

Statement of the Psychophysical Problem: Truth or Fraud of Consciousness

Subjectivity exists. It either is what it claims to be, or it enacts a stage play behind which another type of happening hides. In the first case, its testimony—for example, that I raise my arm because I will it—is credible at face value; in the second case, it is deceptive, a mere disguise or window dressing of neurophysiological processes, which parade in the fancy dress of will but lift the arm without will or the cooperation of will, that is, they do so irrespective of the presence of a "willing" sensation. The standpoint which grants the psyche its effectiveness stays in agreement with its self-testimony and needs no further reasons; the standpoint which denies its claims must have special reasons for doing so. These reasons, whatever their strength, cannot silence the disputed testimony itself, and thus the

natural standpoint, constantly nourished by the subjective evidence, is never abolished in fact. But deprived of the privilege of naiveté once the question of credibility is raised, it must defend itself against those reasons, since at the serious suspicion of an illusion even its irresistibility ceases to count in its favor. On the other hand, the suspicion must indeed be serious. Thus one must first examine the reasons which here contest the validity of immediate evidence and put naiveté in the role of a theatergoer who takes the play on the stage for reality.

Clearly, only the strongest reasons count when it is a question of condemning all mental life to the status of illusion. I can see two such reasons and confine myself to them, ignoring all weaker ones. They are: (1) that any action of mind on matter is *incompatible* with the immanent completeness of physical determination, that is, that the latter does not tolerate such an interference from outside: this I call the "incompatibility argument"; and (2) that the mental as such is also *incapable* of intervention, being nothing but a unilaterally dependent concomitant of physical events and lacking any force of its own: this I call the "epiphenomenon argument." The first argues from the nature of the physical, the second from the nature of the psychical. How strong is either?

I. The Incompatibility Argument

1. *The Argument*

The proposition that the context of physical determination is "closed" and does not tolerate the intrusion of nonphysical causes follows from the rule of the laws of nature, especially the constancy laws, which would be violated each time when a causal quantity with no physical predecessor were added to the given sum or subtracted from it with no physical successor. One or the other would happen whenever in the acting of living subjects the further course of events differs from what it would be without the intervention of the psychical factor, that is, by the corporeal mechanics alone. In the physical reckoning, the intervention would amount to something emerging from nothing or vanishing into nothing, and this is excluded by the constancy rule. Ergo: there cannot be an influence of the mental (nonphysical) upon the physical; ergo: things proceed exclusively according to the physical *concatenatio causarum;* ergo: the addition of the ~~physical~~ psych (subjective) dimension in living beings is gratuitous and redundant for the course of events; ergo: the consciousness of aims, etc. (feelings of willing and acting) is but a deceptive imagery for the causal working of the bodily mechanism—a deception not even excused by a purpose, since *ex hypothesi* the self-sufficiency of the masquerading facts needs no such help: a purposeless deceit of purpose.

2. *Critique*

a) Absolute determinism in physics—an idealization How sound is the argument? We observe that it invokes not simply the validity of the constancy laws, which is to be taken as inductively proven, but their *unconditional* validity, that is, one impervious to exceptions, and this, of course, lies beyond inductive proof. Inviolability on principle pertains to the logical nature of mathematical, not of factual, rules. For the latter, it is merely postulated by us for the sake of the idea of lawfulness. The postulate originates in an idealization and expresses an ideal. The incompatibility, therefore, which the argument states, is of the type that "what must not be, cannot be." The force of the "must not" is proportional to the theoretical dignity of the ideal from which it issues. But since we, and not logical necessity, have invested it with that dignity, we can also reconsider and, if need be, modify it, provided we stay in agreement with observable facts.

b) The logic of the incompatibility problem Moreover, an incompatibility of the nonlogical sort we are faced with designates as such no more than a difficulty of thought and leaves open *what* must be revised for its resolution: the concept of that which is to conform to a norm, or the norm to which it is to conform, or both? To decide this question, we must compare the strength of evidence which both have on their side, but also the *consequences* from any one side yielding to the other: in our case, by asking what becomes of "nature" if her causal purity, or her "exactness," is adulterated, and what of "mind" if its effective power is denied. The scales are tipped by which sacrifice is theoretically more insufferable, that is, more devastating for the side that is to make it. It is a question of the relative price of compatibility, if a price is to be paid. The contention I am going to argue is that on the side of "nature" (where the all-or-nothing logic is inappropriate), the required concession in deterministic rigor would not be devastating for its scientific concept; whereas the concession asked for compatibility's sake from the mental side, namely, the forfeiture of causal force, destroys its concept completely and even drags the favored physical side down into its own ruin, leaving a caricature of "nature" as such.

c) Historical overstatement of the determinist case The seriousness of the problem lies in the challenge of materialist science to inner experience, which has on its side immediate self-certainty but no systematic predictive science; whereas natural science, with no immediate evidence for its generic ideal of objects, can produce constant heuristic confirmation for it in the systematization of phenomena. Because of this tested verification, a conflict with the ideal is a serious problem. But in the psychophysical

incompatibility verdict the ideal is granted more than it is entitled to by its heuristic yield (its only evidence) and more than is necessary to preserve its rightful epistemological position. Unconditional determinism has always pretended a greater knowledge of nature than we possess and ever can possess. The holding of the constancy laws admits of degrees of rigor in detail. There is no *a priori* certainty that what holds for the whole also holds for all its parts down to the smallest; and what holds for the end result holds also for all intermediate links; and what for measurable time intervals holds also for each instant. The identical validity for every part and every instant presupposes an exactitude of nature which precludes any "more or less" and "approximately" and makes nature as pure as mathematics is. It was from mathematics that the idea of such an absolute exactness was borrowed, together with the homogeneity of whole and parts required for its application. *Vis-à-vis* nature it cannot be more than a postulate with which the order-loving human mind perhaps hits the truth of nature, perhaps overtaxes it.

II. The Epiphenomenalist Argument

Leaving the case of "nature" and its alleged causal closure in abeyance, we pass from the argument that the physical does not tolerate psychical interference to the converse one that it need not fear it since the psychical (the "subjective") is devoid of all causal force. The most complete expression of this position is the "epiphenomenon" thesis, whose reasoning runs somewhat like this:

1. *The Argument*

a) The primacy of matter There is matter without mind but not mind without matter. The first is demonstrated by all lifeless nature plus a good part of living nature; the second by the fact that all "mind" (here the blanket title for subjectivity of every kind and degree) appears only in conjunction with certain organizations of matter—organisms, nerves, brains—and that no example of bodyless mind is known. This observation suggests that matter has independent and primary being, mind only secondary being derivative from it.

Experience further teaches that matter in these forms of organization not only provides the necessary, enabling basis or precondition for mind's *existence,* and not only is the originating cause of that existence, but is also the continually determining cause for its *working* and all its changing contents—thus its necessary and sufficient cause in every respect. This is demonstrable for sense perception and certain feelings and emotions, and hence can be extended to thought processes for which it is not yet demonstrated: all these inner, subjective phenomena are wholly the effect

of physical causes. But note that from this causation they do not gain a causality of their own in continuation of the former (as do effects in general), since they are nothing but an expression of what happens in the physical substratum. Mere expression cannot influence what it expresses, nor even itself, since then it would cease to be mere expression. Thus the "induced" subjective appearance cannot emancipate itself in either sense: it can as little play within itself as it can react on the "inducing" substratum. Its pure "as if"–pretense of both, which it displays to itself, has only the entertainment value of an illusion.

b) The completeness of physical determination Joined to the impermeable causal completeness of the physical substratum previously argued, the epiphenomenon status of mind then compels the conclusion that in the case, for example, of raising my arm, *only* the objective, neuromuscular explanation (or description) is correct, while the subjective one—in terms of will and intention—is a nonauthentic symbolic transcription for it. Neurology, communication theory, and cybernetics are at work to implement this postulate—still largely empty with regard to the higher functions of consciousness—by concrete mechanistic explanations or model constructions for increasingly complex cerebral processes. The initial successes justify the expectation, so the protagonists say, that no barrier of principle will stop progress on this road. For the rest, they challenge the "spiritualists" to show how to represent theoretically a moving of physical entities by nonphysical movers and not play havoc in the attempt with the laws of physics.

c) The redundancy of subjective purpose Among the subjective phenomena there are the so-called ends or purposes. For them, too, it must hold that their subjective presence has no possible influence on the course of things. Rather, as already their presence only mirrors an objective state of the substratum, so also the putative acting "*because* of them" ensues in truth *from* the same material conditions of which this appearance itself had been a symbol. The counterwitness of subjectivity on this point is not to be accepted. Socrates, therefore, was not right when, in the famous discourse of his last night, he rejected the corporeal explanation of his sitting there and awaiting the hemlock and proclaimed the explanation in terms of mind—his ideas of right and duty—as the only true one. Neither are those right who view the two alternatives as complementary, as equivalent and interchangeable aspects of the same reality. Only the Ionian naturalists had the right view, for only the physical description explains what physically happens.

d) Mechanical simulation of behavior The test is here, as already Descartes had laid down as a rule, the possibility to *simulate* behavior by

mechanical devices (automata), and this possibility has since been extended right into the mental realm, which Descartes still deemed exempt from such simulation. But then there also holds the further Cartesian principle that what is perfectly imitated is thereby no longer simulated but duplicated, and that then the duplicate discloses the *nature* of the original, to which no other principles of operation must be attributed than those needed for its duplication. Even imperfect imitation, if it is merely a matter of degree, establishes the perfect imitation as theoretically possible. Thus, if intelligent purposive behavior *can* be simulated in some simple forms, the more complex ones are *in principle* covered by the feat. Then the *ideas* of purpose, and other subjective data, which in the original concur with the behavior, are redundant for its actual performance and thus have no role in it. Thus the newly discovered possibility of imitating mind by purely corporeal means strengthens the speculative hypothesis of epiphenomenalism about the impotence of mind in general and the functional otiosity of psychological purpose in particular. Like all consciousness which attributes authorship to itself, teleological belief has the character of a purely putative, operatively redundant and inexplicable "as if."

2. *Internal Critique of the Concept of "Epiphenomenon"*

So far the materialist-epiphenomenalist argument. My discussion will focus on the *causal* aspect, which is the hub of the argument, and in its spirit I open the debate with an incompatibility argument of my own. Epiphenomenalism makes matter the cause of mind and mind the cause of nothing. But causal zero-value is compatible with nothing adhering to matter; and in particular it runs plainly counter to the idea of causal dependency itself that something dependent should be an end only (effect only) and not also in its turn a beginning (a cause) in the chain of determination.

a) The first ontological riddle: creation of soul from nothing Before pursuing this line of thought, let us take a closer look at the concept of "epiphenomenon" as such. It says in general that the subjective or the psychical or the mental is the concomitant of certain physical occurrences in brains. The "concomitance" is one-sided, not reciprocal: the physical processes, as the primary reality, are autonomous; their secondary psychical expression is totally heteronomous or a mere product of an other. The presence of the product makes no difference to the history of the producing "other"; neither does the feat of production itself, which as it were happens "behind its back," on the basis of it but without a contribution from it. The product is a *by*-product of the intraphysical producing, with no expenditure deflected to its production, which thus is not a transitive act of the physical base but merely the "appearance" of its immanent func-

tioning. This functioning goes on as it would anyway, whether or not it had occasioned such an accompaniment. Thus the occasioning of it is *causaliter* a "creation from nothing," since nothing was causally spent on it. Otherwise, the epiphenomenon hypothesis were useless, since the quantity expended would physically have disappeared and its converted psychical succession were no longer quantifiable—precisely what was to be avoided for the sake of the constancy principle. *The soul's causation from nothing is the first ontological riddle which the epiphenomenon theorem braves in deference to physics, in which otherwise never a thing is supposed to arise from nothing.*

b) The second ontological riddle: a noneffective physical effect What thus has been produced from nothing must also remain a nothing, causally speaking. Just as the occasioning of the "accompaniment" must cost the occasion nothing, so also its being there must not change the happening which it accompanies. This indeed is the first concern of the epiphenomenon thesis: the monocausality of matter, to be safe from interference, demands the *impotence* of mind in the first place, then its cost-free generation in the second, if it is also to come *from* matter. Admittedly, the construction becomes at least logically consistent thereby: only that which is made from a causal null can also remain a causal null, having inherited no force from its cause. And yet it *is* something itself and not nothing, its being there clearly different from its not being there. The appearing of consciousness adds something to the composition of reality by which it becomes different—descriptively, but not dynamically: it "is," but nothing follows from it for the rest of things, nothing propagates from it into their further course—the rest of reality remains what it would also be without this mirage in its midst. That which is endowed with consciousness does not behave as it does because it is conscious but is conscious because it behaves as it does, that is, as its physical makeup makes it behave. Accordingly, the concept of consciousness is negatively determined thus: a something for itself and yet a nothing for other things, a being without consequence, a noneffective fact. *That something, itself a consequence (an effect), be barren of consequences, is the second ontological riddle which the epiphenomenon theorem braves in deference to physics, in which otherwise nothing is supposed to remain without consequences.*

c) The metaphysical riddle: existence of a "delusion in itself" The no-consequence rule applies to the mental in two directions: outward toward the physical sphere, as just explained, but also inwardly toward its own continuation. That is, the impotence must, besides that of determining the body in action, also include that of self-determination of thinking in thought. Otherwise, mind would cease to be epiphenomenon and could

go its own ways, on which the concordance with the body events might be lost. In that case, the illusion would explode when the mental sequence reaches the point of action, that is, of intentional determination of the body, and the impotence would be revealed: I will one thing and my arm does the other. But it is precisely the appearance of *power* which the impotence thesis is designed to "save." True appearance of the impotence would falsify the theory of impotence, which is just a theory of the deceit and not of the truth of consciousness. It is essential for the impotence thesis that the impotence remain hidden behind the appearance of power. To ensure this, the impotence must be indivisible: only with equal inward impotence can the outward impotence remain hidden. Internal power with external impotence would result in a consciousness at odds with the world. But the theory was devised to show a consciousness united with the world—first of all with the body. Since only the world, namely, matter, has genuine reality, the unity is nothing but the negation of *any* self-reality of the other side, therefore of internal as well as external causality on its part.

This is just what the concept of epiphenomenon provides for, by denoting something which each moment originates anew from the basis, and whose continuation, therefore, is not its own but that of the basis. As, in a motion picture, the next phase of a movement seen on the screen does not originate from the preceding phase, as it appears to do, but independent of it from the projector that emits both, and thus on the screen, contrary to appearance, in truth nothing moves—so also the temporal successor of a "now" of consciousness cannot come from this but like itself must come from the physical substratum of which each state of consciousness is, by definition, the epiphenomenon. Thus it "mirrors" the progress of the substratum, while appearing as progress of itself. As in the movie, this appearance is mere illusion. The hammer shown to crash down on the anvil is an image sequence, not a dynamic action; the deliberation terminating in a resolution is a sign sequence, not a real bringing-about: its dynamics too is illusory (the internal one, even before its external sequel when, e.g., as a seeming result of the inner sequence, a real hammer is swung).

The real dynamics which the sign sequence represents is the cerebral one, which expresses (and at the same time conceals) itself in the "text" written by it on the page of consciousness; but the continuity of the text is nothing but the continuity of the writing of it, which from below ever anew generates each sign of which it is composed and by no means lets the text write itself. The way goes not from sign to sign but from brain state to brain state and only hence each time to its equivalent in the sign script. To stay with the movie comparison: it is the continued projection that generates the continuation of the projected image; this itself is powerless. So also the "sign" on the "screen" of subjectivity in relation to

its successor. No thought engenders further thought, no state of mind is pregnant with the next. But since precisely this is intrinsically claimed by them, it follows that all thinking is deception already in itself, and then once more when it believes to pass outside itself into bodily action. The delusion of self-determination overarches the delusion of body-determination. Both delusions are essential to the psyche. "Soul" *is* that very delusion or make-believe with which reality constantly deceives itself. Itself? But no, brains are not deceived. The subject? But this already is a deception as such. And why? Again no answer, not even that of *l'art pour l'art,* which fits Descartes' evil spirit, but not unintentionally fraudulent matter. The deceiver, whoever be the deceived, earns nothing from his deceit. The game is played to no player's benefit and no victim's harm. It cannot serve a purpose as such. "Purpose" is a cipher of something that is by nature purposeless. The script in which the cipher occurs—mentality—has itself no purpose, let alone a function. Its pageantry as a whole is a mirage. *The existence of such a "delusion in itself" is the absolute metaphysical riddle which the epiphenomenon theorem braves in deference to physics.*

There are other riddles, of a more logical kind, to be exposed in the concept of epiphenomenon, but those shown here may suffice. Riddles by themselves do not dispose of a theory, but they surely make it suspect. We now advance some arguments against it: (1) its internal inconsistency in that it violates the very concept of nature to which it pays homage, (2) its absurdity as a theory-destroying theory.

d) The causal inconsistency of the concept The inconsistency was mentioned before: "soul" has to be ineffectual so that causal law be inviolate. But from this same law the "epiphenomenon" itself would be a singular and inexplicable exception—in coming to be at no causal cost and in existing with no causal effect. Changing roles, *we* must now defend the principle against its defenders and insist that *nothing* in the world is "for free" and nothing, once there, ends with itself, that is, leaves no issue in the world. Everything caused must itself become a cause. But the "soul," so materialism holds, belongs to the world as a result of matter. Then (*a*) matter must spend something on it, and the balance sheet of a material process must read differently, depending on whether or not it has engendered an effect in the form of consciousness. *Something* of it *must* have passed over into the effect, even if in this case we do not know the equivalences for an accounting. Something was exchanged for something. Conversely, (*b*) the existence of this new datum (the resultant copresence of mind), surely different from its nonexistence, must have a share in the progress of events, since no difference of existence is dynamically neutral—even if here again we are ignorant of the manner of transfer. Only

on this condition has the expenditure for its becoming not simply vanished from the world, and the balance of the whole is preserved.

The principle is simply that as little as something can arise from nothing it can vanish into nothing. Nothing is pure beginning and nothing pure ending. Only the indeterminist can admit exceptions from the rule; but the materialist argues the cause of determinism and is bound to it. Thus his "solution" contains a real self-contradiction, namely, saving the inviolability of a principle by its violation.

3. *Reductio ad Absurdum from the Consequences*

So much for the internal critique of the *concept* of epiphenomenon. More devastating still are the *consequences* which flow from it for everything else: for the concept of a reality that indulges in this kind of thing, for a thinking that explains itself by it, and for itself as a thought of that thinking. Here the charge is not inconsistency but absurdity.

a) An absurd nature First, what sort of being would that be which brings forth, as its most elaborate performance, this vain mirage? We answer: not a merely indifferent, but a positively absurd or perverse being, and therefore unbelievable. If living behavior were nothing but a deaf-mute pantomime, performed by supremely sophisticated physical systems without enjoyment of subjectivity, it could well be termed pointless but not strictly absurd. The show becomes absurd when it accompanies itself with a music *as if* its predecided paces were set by it. A lie can have a function, but not here: the mechanical needs no bribe. And yet it should sound—in will, pleasure, and pain—a siren song with no one there to seduce? A song that only sings its error to itself, including the error of being the singer? Something devoid of interest in the first place, and with no room for its intercession in the second, should stage the grandiose comedy of interest, shamming a task that is not there and a power it does not have? The sheer, senseless futility of such an elaborate hoax is enough to disqualify it as a caricature of nature. He who makes nature absurd in order to circumvent one of her riddles has passed sentence on himself and not on her and has forfeited the right to speak anymore of laws of nature.

b) A theory-destroying theory Even more directly than via the slander of nature has he passed judgment on himself by what his thesis says about the possible validity of any thesis whatsoever and, therefore, about the validity claim of his own. Every theory, be it the most mistaken, is a tribute to the power of thought, to which in the very meaning of the theorizing act it is allowed that it can rise above the power of extramental determinations, that it can judge freely on what is given in the field of representations, that it is, first of all, capable of the *resolve* for truth, that

is, the resolve to follow the guidance of insight and not the drift of fancies. But epiphenomenalism contends the impotence of thinking and therewith its *own* inability to be independent theory. Indeed, even the extreme materialist must exempt himself *qua* thinker, so that extreme materialism as a doctrine can be possible. But while even the Cretan who declares all Cretans to be liars can add, "except myself at this moment," the epiphenomenalist who has defined the *nature* of thought cannot make this addition, because he too is swallowed up in the abyss of his universal verdict.

Thus we have a twofold *reductio ad absurdum,* according to the twofold question of what to think of a reality that brings forth this futile image, and what of the attempt of this self-confessed mirage to establish a truth about that reality. Nature as an impostor on the one hand, a theory destroying itself on the other, was the outcome of the scrutiny.

III. "Epiphenomenalism" Voided by the Voiding of "Incompatibility"

But, so we ask at last, was this suicidal *tour de force* really necessary? Are the reasons prompting it indeed so compelling that they only leave this counsel of despair—or that of conceding defeat before an insoluble riddle? The latter, of course, would be the proper option if this were the choice, and there would be nothing shameful about its modesty and honesty. But I think one can do better. I think one can show to the upholders of physical causality that their cause admits a resolution of the dilemma that does not injure it on their own terms. To this end I take now the liberty to engage in a thought experiment which is not meant to produce a theory but merely to illustrate possible compatibility by a freely constructed logical model.

Tentative Solution of the Psychophysical Problem

1. *A Thought Experiment*

Let us assume that a geometrically perfect cone stands with its apex on a surface over a center of gravity which lies exactly in the prolongation of its axis, that is, this rises exactly "vertical" from its support. With perfect symmetry and complete absence of other forces or force differentials the cone would stand in absolute, but absolutely unstable, equilibrium. "Absolutely unstable" means that an infinitesimal influence would suffice to tip it over and make it fall to one particular side. Let us think it to be a giant cone: the smallest disturbance of symmetry from without or within would trigger giant consequences. These consequences in their total course, in acceleration, force of impact, mechanical and thermal effects, equalization of the initial energy gradient, would each and all be governed univocally by the laws of nature and be calculable according to

them—except the accident of the direction of the whole, for which all other directions, that is, an infinity of such, were equal candidates. As to the trigger impulse that did decide the direction, its minimal value does not enter into the measurement of the consequences with their incomparably higher order of magnitude, and for purposes of *their* causal description the unknown starter x equals zero. To be sure, the universal physical hypothesis would postulate for this x again, infinitesimal as it be, that it was determined in accordance with the constancy laws; or generally that there was a "sufficient reason" in the antecedents for this one direction of tilting being selected over all the equally possible ones. But operationally, the postulate has to remain empty for lack of verifiability, and it is obvious that here, at the critical balance point, there is room for sheer randomness or indeterminacy without detriment to the strict determinacy of the movement once underway. Indeed, computationally it would make no difference if for the initial x we posited a psychic, nonphysical origin. In the present case there is no reason whatever for doing so, but this may change as we come to other examples. So far the thought experiment shows that neutral interstices in the matrix of physical determination, properly located, would not interfere with the validity of physical law and the intactness of causal "bookkeeping."

2. *The Trigger Principle in Efferent Nerve Paths*

Our first example was nothing but the simplest, crudest, and thus utterly unrealistic illustration of the *trigger principle,* which plays such a crucial role in higher organizations of matter, that is, in the kingdom of life. Let us go straight to the summit of the pyramid where we encounter the subtlest form of the "inverted cone." Assume we have the trigger points *A, B, C* . . . , at the primary control centers of efferent nerve paths in the brain, corresponding to the respective possible motor commands *a, b, c* . . . , thus representing the "yes or no" for the actions $\alpha, \beta, \gamma \ldots$; and further assume a physical state in which the chances of activation are equal but alternative for all (any one, but one only, being eligible), that is, that the decision on *which* of them will be "fired" is completely in the balance; and finally assume that for the activation, that is, for occasioning the transition from potentiality to actuality ("triggering"), an influence of the smallest order is required. What then would be the situation physically if the "choice" has actually fallen on *A?* Well, the ensuing transaction *a,* that is, the neural tranmission of the "command," and then the transaction α, that is, its motor execution (and thereafter everything which follows from it in the external world), can in unbroken sequence of determination be described, and thus explained, according to the laws of nature (thus ideally also be predicted)—without there having to be an answer to the question of why *A* rather than *B* or *C* had been activated. Since the magnitude responsible for that has zero value for the

account of what is observable thereupon, it makes no difference for the latter's consonance with the laws of nature whether *A* or *B* or *C* has been activated. The alternatives are equally orthodox in their physical behavior, equally possible "before" and equally deterministic "after," and only the decision on which of them is allowed to become operative is indeterministic *on this plane* of calculability.

Nevertheless, for this *x* too (i.e., for the initial resolution of indifference) one might in theory at least insist on a physical accounting, considering that the relative "zero" of the triggering factor is, of course, not a real zero but *some* quantity (even, according to quantum theory, of an indivisible minimum value), and one has a right to ask whence this came and what *its* prior determination was. Two answers are possible on physical terms: either its incidence on *A* was a purely random event in the given quantum field, or it followed deterministically from the antecedent distribution therein. Both answers are physically acceptable on principle, but both here would be in contradiction with the facts: the randomness with both the physical and mental facts, the mechanical determinism with the mental alone (as shown before), in that it would allow them no power at all.

3. *Possibility for the "Triggering" to Originate in Mind*

There is a third alternative, which admittedly would no longer be a physical explanation: that the physical quantity required for the selection of the neuron (a sort of "Maxwell's demon") is *generated* on the part of "subjectivity" or "psyche," that is, from beyond matter. Let us not be too frightened by the anathema of this idea to physical orthodoxy to examine its implications. It speaks of a *de novo* increment to the antecedent physical sum, the insertion of a value not accounted for from within the physical system, and in that respect a *creatio ex nihilo*. Its smallness would make it unspottable in any computing of physical factors for the macrosequence of events, but the effects of its trigger function in the hypersensitive balance of the threshold condition can be immense: war and peace, rise and fall of empires, building of cathedrals and atom bombs, greening or wasting of the world may be "caused" by its infinitesimal contribution. All these chains of action in their macrocourse would offer to causal analysis a deterministic picture in complete accord with the constancy laws, as *also would innumerable others:* only the prebeginning of each, which selected it from among the alternatives, would be indeterministic, that is, not physically *but mentally* determined—just *as direct experience tells us*. From the point of view of physics, the one "direction" actually taken would be (as with the cone) an absolute accident, which as such is beyond verification.

So far, so good. But an objection is obvious: many vanishing magnitudes can add up to noticeable ones; and what we have been speaking about

was not a rare event, but something occurring constantly and in countless cases, namely, with every outward action of every "subjectivity." Thus, even if willing, uneasily enough, to grant the occasional happening of a sufficiently small "creation from nothing" in the context of physics, we may well balk at what our suggestion seems to amount to: that the world of life, across the whole breadth of subjectivity diffused through it, incessantly donates antientropic energy to nature, unpaid for by increased entropy elsewhere. This, apart from its repugnance to general theory, robs the hypothesis of its saving grace of being innocuous to the rule of the constancy laws, since the accumulation of the singly nonmeasurable must eventually grow into the dimension of the measurable and, hence, into visible conflict with the requirements of those laws.

4. *The Dual, Passive-Active Nature of "Subjectivity"*

So it would indeed, if the case as presented up to now were complete. But its reverse has yet to be added, and when complemented by it, the hypothesis by no means stipulates a one-way flux of becoming. First we simply recall that matter was supposed to be the prior donor in supporting subjectivity as such, and so it may just receive back its original outlay in the "influx" we discussed. More especially we must remember that—as afferent nerves correspond to efferent ones—consciousness in its overall world-relation is essentially a two-way street and not a one-way street. Action *into* the world, thus far considered alone, is based on information, an input *from* the world, that is, ultimately on sensibility. But in every act of sensuous affection, the physical chain terminates in a mental representation, namely, the percept in question, and this too cannot be free of causal cost: some value must have vanished from the physical ("objective") side to reappear on the mental side in the radically different form of subjectivity. If we were nothing but contemplative beings who only perceive the world, there would be constant transfer and drain from the physical order, thus the same embarrassment as before with the opposite sign; if only active beings (e.g., with an *a priori* intellectual knowledge of all objects of action), then there would be constant reverse transfer and inflow with *its* embarrassment. Since we are inseparably both, and this in essential complementarity, it is not unreasonable to assume that, in the physical average, outflow and inflow balance each other with reference to the whole phenomenon of subjectivity, and the two opposite embarrassments cancel out.

The key to a solution of the psychophysical problem, that peculiar impasse which nothing but a philosophizing physics has created for theory, lies—so we suggest—in an age-old insight which has never been utilized in this connection: that our being *qua* subjects has this double aspect and consists of receptivity and spontaneity, sensibility and understanding, feeling and willing, suffering and acting—in brief: that it is passive and

active in one. To what model construct, then, does our thought experiment lead?

5. *The Speculative Model*

Using a metaphor, we say that the net of causality is widemeshed enough to let certain fish slip in and out. Or with a change of metaphor, at the "edge" of the physical dimension, marked by such peaks of organization as brains, there is a porous wall, beyond which lies another dimension and through which an osmosis takes place in both directions, with a priority of that from the physical side. What thus physically seeps out and in is of too small a magnitude to show up quantifiably in the single case and mutually so balancing in the total as not to affect the verifiable overall working of the constancy laws. In virtue of the trigger principle, the smallness of the single input or output does not preclude great physical effects. Passage through the "wall" means each time a radical transformation in kind, such that any relation of equivalency, even the very meaning of quantitative correspondence, ceases to apply. The greatest thoughts with the mightiest consequences can arise from the tiniest physical input, and the tritest just as well. What matters is that between input and output there is interposed a process of an entirely different order from the physical one. Short or long as may be the loop of the circle that passes through the mental field on the other side of the wall, it does not move by the rules of quantitative causality but by those of mental significance. "Determined" it is too, of course, but by meaning, understanding, interest, and value—in brief, according to laws of "intentionality," and this is what we mean by freedom. Its yield is eventually fed back into the physical sphere, where everybody can recognize it (for everybody knows that unthinking nature builds no cities), without any single physical nexus confessing to its share. With the transfer forth and back, where egress and ingress go on continually, the total balance for the physical side remains even (nothing analogous applies to the mental side), and it is on the plane of that balance that natural science does its explaining. The *understanding* of the same event is done from the plane of that which for the moment stands outside the balance and is "transcendent" to it in this sense. In that understanding, the extraphysical interlude is recognized as the true origin of the physical action, though only infinitesimally its "cause."

For mentality, thus, the brief formula holds: generated by minima of energy, it also can regenerate minima of energy. In between, these minima are gone from the physical surface, yet have not vanished into real nothing anymore than has a subterranean river; thus, they also do not emerge from a real nothing when consciousness acts back into the world. The "in-between" itself is the realm of subjectivity and (relative) freedom.

Accordingly, the brain is an organ of freedom, but precisely on condition

that it is an organ of subjectivity. To put it the other way around: supposing a brain of the same physical constitution as the human, but without concomitant subjectivity—we contend that it would not produce the same effects in the visible world (though perhaps quite respectable ones in body control) which we know the human brain to do. That is to say, subjectivity is not a causal superfluity. No more need to be said here about the difference in the "intelligence" of machines and of the human mind. Actually it is my view that the hypothesis of a "merely physical brain" (except in a corpse) is inadmissible. Such a physical organization *eo ipso* means in its functioning the opening up and sustaining of a psychic dimension, which then participates in the overall causality of the system with the leverage of its key position. Obviously the freedom thus established is not absolute but confined to the latitude which physical necessity itself allows it. It does allow, as we have illustrated, that the smallest force can wield the greatest power when in the given "critical" configuration it suffices to "tip the cone."[3]

That there are, in recurrent readiness, situational sets of such poised "cones" and thereby sets of options among physically equivalent possibilities is the functional meaning of such physical organizations as brains in control of organisms. Riding on the crest of this physical organization—one of whose roles is that of amplifier—"mind" with its immeasurably small physical input can be the initiator and determinant of physical effects in that order of magnitude in which visible behavior takes place.

Ontologically it should be noted that, according to this model, the "beyond" of the dividing wall or membrane is not a no-man's-land which keeps its inhabitants for itself and in which they can lose themselves as in a spirit world. Just as it only lives on the continuous input from the physical side, so it feeds back into it what has gone through the transformation of subjectivity. Mind thus belongs to the one and same ontic reality as matter, only with a thoroughly different nexus of ontologically different elements within its own dimension. In other words—as the voice of the "self" has always been telling us—the one, coherent, convertible, and intercommunicating Being is not exhausted by its massively prevalent physical aspect.

6. *Evaluation of the Model*

The model we have constructed is admittedly crude. We need not bother with whatever refinements it is susceptible of, since it does not even claim to be "true," that is, to portray what is actually the case. It is a mere play of thought, meant to illustrate that *on the "physicalist" premises themselves* the psychophysical impasse of their making (and their alone) logically admits of better solutions than the wholly unacceptable one of epiphenomenalism. As a point of strategy, not of conviction, we have made maximum concession to the materialist case and its conception of

the nature of matter. The truth, I suspect, would look vastly different—not only more subtle but also framed in ontological terms which would alter our very speech of "matter" and "mind."[4] For our purpose it was enough to come up with a hypothetical fiction in the conventional terms that satisfies these three requirements: to be self-consistent, to be consistent with observable facts, and to spare nature and ourselves the scandals which materialist epiphenomenalism has been shown to inflict on both—on nature the scandal of an alleged procreation of mind without causal cost and consequence, on mind the scandal of utter futility in both action and thought. We have at least recovered, even for the captives of physicalist creed, a good intellectual conscience for believing in the immediate evidence of our being.

IV. Quantum-Mechanical Review of the Proposed Solution

The first publication of the "tentative solution" to the psychophysical problem, in 1976, had one highly welcome result in that it enticed Professor Kurt Friedrichs of the Courant Institute of Mathematical Sciences of New York University to volunteer his critique and constructive comment. In many hours of discussion over repeated sessions, he reviewed the issue and explained that a "model" answering it—if any can be found—must be framed in terms of quantum theory and not of classical physics. I shall try to communicate the instruction thus received as it bears on the thought experiment I had initially proposed. The shared premise was that "epiphenomenalism" is indeed untenable, and a better alternative to it must be found. Professor Friedrichs's observations were meant to help the search for it. I gratefully acknowledge his much-needed and so generously given aid. Needless to say, the responsibility for what follows is still mine—a layman in these matters.

1. *The Incompatibility Argument: Valid on the Plane of Classical Physics*

As to the present inadequacy of the model, the objection was not so much to the inelegant idea of the mind swallowing and regurgitating physical magnitudes that have meanwhile vanished from the physical scene, with the (ever so slightly and momentary) blurring of the laws of classical physics this entails, as to the attempt in general to wrest any latitude for an indeterminacy of any kind from classical mechanics, to which my model remained conceptually wedded. Prequantum physics, it was stressed, cannot deal with the problem of mind except in a completely determinist manner and is thus truly not compatible with an interaction between mind and physical processes, if mind is something other than physical process. Of particular relevance to our problem is the principle of action and counteraction (force-counterforce), without which no causal transaction

is conceivable. But the subjective sense datum, for example (as distinct from the objective physiology of sensing), granted its being in some sense an "effect" of a physical cause, does not in turn *react* on it, or together with it constitute a case of "counteraction," and in the incommensurability of the mental and physical as such, a dynamic correspondence of that kind (whichever way the traffic goes) has simply no place. There is no real interaction. In sum: prior to the advent of quantum physics, the incompatibility argument was right.

2. *Indivisibility of Inner and Outer Power of Subjectivity*

With regard to the *selection* of alternative physical possibilities by mental choice, note should be taken of two different meanings thereof: (1) purely mental: sitting in my armchair, I select Carter as my candidate, that is, I "make up my mind" without at this time performing an "action"; (2) I elect Carter by pulling a lever in the voting booth, that is, out of my mental determination I perform a *physical action*. The manifest locus for the psychophysical problem is (2), and for the physicist the only one, though it was shown in the original argument (II.2.c) in this Appendix, that already (1), that is, the self-determination of mind in thought, does raise the problem with its claim to autonomy *vis-à-vis* the physical realm, and that ultimately "power" or "impotence" of mind holds either in both respects or in neither. Considering that thought itself has, or depends on, a physical organ, the brain, one can even argue that the two cases differ only as intracerebral activity and brain-transcending motor activity, both being equally physical processes, one on the microscale, the other on a macroscale. But we are not aware of the first, and only the second involves visible and willed changes in the external world: and it is here, in the transition from volition to macroscopic action, where the trigger principle proves helpful in explaining something of the *outward* path, at least, of the psychophysical dynamics. (The opposite path, as great a riddle, was not further discussed.)

3. *Aspects of Quantum Theory*

Passing over to quantum mechanics, particular emphasis must be put on this tenet of the theory: it is not possible to know the state of a physical system to such a degree of completeness that its future states can be univocally predicted. The barrier is fundamental and not in the nature of a temporary halt in the perfection of observational techniques. It is indeed an integral part of the theory itself.[5] Familiar expressions like "state description," "determined," "prediction," assume an altered meaning. There can indeed be a description (called the Schrödinger function ψ) of the state of the system such that, if it is known for a particular time t_0, it is also known for any future $t' > t_0$—but only if no new measurements are made. Even if new measurements are made, probabilities for their

outcome are determined by the Schrödinger function obtained from the original measurement. This sounds "classical" enough. However, the information thus provided is not what classical physics would have called a complete description of the state of the system for all times $t' > t_0$ if the state at t_0 is known. It merely allows us to compute probabilities for the outcome of certain measurements (or of certain interactions of the system with other systems) in the future. This is what "state of the system at t_0" *means* in the context of quantum theory. In some cases, the possible outcomes of measurement are very sharply defined, and thus we can even say that they will assume, with great precision, only particular values, for example, the future energy state of an atom will predictably be one of the values $A, B, C, \ldots$, but none in between. Among the eligibles themselves the choice for coming out a winner is open and only subject to probability gradients. Thus prediction, derived from ψ, of the outcome of a measurement at $t' > t_0$ is *not univocally* defined. If we still wish to speak here of causality, we must note that this concept differs now from causality in classical mechanics in two striking respects. First, it is on principle impossible to measure simultaneously the values of all the quantities connected with the system, such as position, velocity, energy, etc., and no set of future measurements can recapture all of the values of the other parameters at the moment of the first measurement.[6] Second, in general, any sharp measurement changes the "state" (in the sense defined) of the system. What one can measure—to any degree of exactness—each time is *one* of certain complete sets of compatible quantities that are said to determine the "state" or the Schrödinger function of the system at the given moment. ("Compatible" means: simultaneously measurable without mutual interference.) Then the "state" defined in this narrow sense is indeed determined in the future—but if, at any later time, one measures one of the incompatible quantities, the state "collapses" and is replaced by a new state description. Also, the sharper the values of one set of compatible quantities are measured, the less sharply known are the simultaneous values of the alternative set: from a middle ground of equal inexactitude for both, the ratios of simultaneous knowledge for the incompatible quantities stretch to the opposite extremes of all—zero, where the product of the indeterminacies is constant (Heisenberg). In sum, at no one time is the state of the system, and therewith the outcome of future measurements, completely knowable; it is, in this sense, an "ultraphysical"[7] reality (Friedrichs).

4. *Possible Uses of Quantum Theory for the Psychophysical Problem*

The bearing of all this on our subject is that here, at a crucial spot in the texture of things, a gap parts the web of our knowledge, a gap not given to being closed on principle, but sufficiently defining the *terra incognita* within for us to understand that in it we can no longer appeal to a deterministic physical theory.

The attempt to locate *in this "gap"* a solution to the psychophysical problem, or to that of "freedom and necessity" (or there to seek an *asylum ignorantiae* for it), is not new.[8] The tendency appeared almost from the beginnings of quantum mechanics and principally fastened (as far as I am aware) on two aspects of the theory: on the principle of complementarity and/or on the principle of indeterminacy.

a) Complementarity Niels Bohr himself, who in his philosophical moods was much inclined to extend the formalism of his complementarity principle to areas beyond its native soil, at one time tentatively suggested its possible application to the psychophysical problem,[9] never (to my knowledge) to take it up again. But the wide currency which "complementarity" has gained, not without Bohr's encouragement, as a generalized logical tool for "dual aspect" situations of all kinds outside its home ground (notoriously so in the social sciences), places it in the field as a candidate for tackling the primordial "dual aspect" case of human experience, the psychophysical. I regard all these extramural uses of "complementarity" as highly dubious and at best metaphorical; for the psychophysical problem I will try to show that it logically does not fit the situation at all.

First a few words about the original, quantum-mechanical sense of "complementarity" as coined by Bohr. It concerned the possibility of defining the "state" of a system in terms of two, mutually exclusive, conceptual representations. One may say that complementarity, in the sense of Bohr, means that we can ask one of two mutually exclusive sets of questions about a system, but not both. The answer to one question would describe the system as a particle; the answer to the other would describe it as a wave. But the two models are not just optional alternatives equivalent with one another. They stand for different observables. For example, the position component X_1 is complementary to the impulse component V_1, thus the particle description answering to the measure of X_1 is complementary to, but not interchangeable with, the wave description answering to the measurement of V_1. The knowledge of one of them precludes the full determination of the other, and the more I know about the one, the less I know about the other; yet both are required for a full account of the phenomenon, complementing one another in conveying its truth, that is, the exhaustive knowledge of what is knowable about it. As quantum physicists are wont to say: The entity "is" a particle when I measure its position, and it "is" a wave when I measure its momentum. Thus, whatever it is "in itself," only a dual account can do justice to the object (or, express the "truth" about it), without therefore bespeaking a dual nature of things.

Now it is tempting to think that something similar might apply also to

the twofold account of human action (and of conscious behavior in general), the "outer" and the "inner," and provide a solution to the ancient problem of necessity and freedom: descriptions of one and the same train of events in terms of physical necessity, on the one hand, and in terms of mental spontaneity, on the other, are "complementary" in the sense of Bohr's principle; the either-or is one of representation, not of fact, and only both representations together *in* their difference convey the truth of the identical fact. As in the case of the particle and wave descriptions, both are equally genuine—and, we should add, equally symbolic, unless we are prepared by analogy to apply to this situation too the boldly idealistic assertion of certain quantum physicists that the "object" itself actually changes, with their measurings of it, from "being" a particle to "being" a wave, or vice versa (which I doubt even they would mean to hold for the corresponding terms in the mind-body "complementarity").[10] Somehow, at any rate, the underlying reality thus doubly expressed is supposed to be one in itself.

To discuss this appealing suggestion thoroughly is not appropriate here. It would turn out to "work" as little as any variety of "psychophysical parallelism," Spinozist or other. One little observation, however, may suffice to show that the transfer is in this case logically at fault from the outset, namely, not faithful to the formalism of the original concept. To "complementarity" in its quantum physical conception by Bohr himself, it is essential that the two descriptions are cleanly separate, each complete in itself and neither intruding into the other: the wave description is not to be contaminated by corpuscular terms, and vice versa. The two models are, in short, strictly alternative. But we cannot begin to describe anything "mental" without referring to the "physical," the world of objects with which mind, sense, will, action have actively and passively to do. That is to say, any speech about mind *must* also speak of body and matter. And when speaking of ourselves, we not only can and do but always *must* embrace with it our physical and mental being *at once:* precisely that *simultaneous* entertaining of *both* sides which quantum theory rules out for its complementary alternatives. From this original, *joint* givenness, after all, the psychophysical problem arises in the first place. Here, the *isolation* of the two components is an artifact of abstraction, their interlocking copresence being the primary datum. Even in abstraction, we noted, the isolation does not really succeed, as the description of one side intrinsically refers to the other. The lines themselves do not run parallel but cross.[11] On this *transitive* relatedness alone, by which one description draws into itself elements of the other, the purported analogy with the quantum-mechanical situation breaks down.[12]

More to the point of our discourse than this formal observation is the blunt reminder that what is substantively at issue in the psychophysical problem is *interaction* and, more particularly (surely so for Bohr's interest

in "freedom"), the question of an *intervention* of mind in the affairs of matter. This, if it takes place at all (which is just in question), is nothing like an invariable concomitant, an innocent complement of physical processes, but is a particular event affecting their course. Does this happen? Is it possible? How? Such a question is evidently meaningless in the case of the complementary wave and particle description—for example, to ask whether, to what extent, and on what occasions the wave aspect of events leaves its mark on their particle aspect. But just those questions (with the appropriately substituted terms) are the most meaningful ones to be asked in the psychophysical setting.[13] Complementarity, "noninterventionist" by its formal nature, does not even allow them to be asked when seriously held to apply. In sum, I believe, no faithful analogue of complementarity as understood by Niels Bohr really applies to our problem, and philosophers should leave it where it belongs. The philosophical interest of its attempted enlistment lies in what it has in common with Spinoza's parallelism of attributes, namely, the *noninteractional* premise, of which it seems to offer a more sophisticated version. But it shares with it the fatal weakness of every parallelism: to leave the body in the dominant position, all assurances to the contrary notwithstanding. For only the body's internal determinism, that is, that of material nature in general, is *known* with the force of predictability, while that of the mind is not. By the logic of theory as such, the "hard" side thus distinguished will always prove the stronger partner in a marriage with a nondeterminist, "soft" mate, who can only play second fiddle to the other's leading tune. Matter will dominate and mind has to follow suit. This is the demonstrable fate of every form of psychophysical parallelism, of which "complementarity" is a variant.[14] "Epiphenomenalism" lurks in all of them. Thus, quite apart from the formal defects, complementarity, in the way it materially predecides the crucial issue, has taken the sting out of the problem that should not let us come to rest. Or, more likely, it is nonrigorously applied, and in that vagueness it is worthless.

b) Indeterminacy A better promise is held out by the principle of *indeterminacy*. To enlist it for the theoretical cause of "freedom," seeing that this is somehow ranged against "determinism," is an obvious choice, and it appeared early in the literary wake of quantum mechanics. Almost as early, however, two grave objections were raised against the idea. First, the working of mind—thinking, making decisions, etc.—even if not determined by physical necessity, is anything but indeterminate; its freedom goes together with high-grade orderliness, and so no principle of randomness can profit a theory of that freedom.

Second, no contribution from that quarter could make itself felt on the plane of macroevents where human acting takes place: there, the latitude of probabilities obtaining on the subatomic plane, where quantum me-

chanics governs, is superseded by the tight determinism of classical mechanics, as with the statistics of great numbers the probabilities from the microsphere turn into certainties in the macrosphere. So, in that respect too, the indeterminacy principle offers no comfort to freedom in that sphere, and none other is worth talking about.

The first of these objections, that of randomness versus the orderliness of mentality, can only be countered by hypothesizing that mind has somehow the power to bend indeterminacy to its purpose—to pick out, as it were, the winner from among the spectrum of probabilities. How to represent, by a model conception, the manner of such an intervention or influence is a matter for speculation and perhaps for despair; but to try for it in the open spaces outside verifiability is surely not interdicted by any veto of natural science or of rational thought in general. Even failing there (as may be fated), the hypothesis as such would not *clash* with the data and rules of that physical locus and may as well enjoy the licence of the *terra incognita* (incurring its dangers too). To this, we shall return at the end.

The second objection, that of the statistical submersion of subatomic indeterminacy, can be more positively met if a way can be shown on which a single quantum-mechanical event may become a determinant for events on the macroscale of our experience and action. Now the *trigger* principle offers just such a way, and herewith I resume the digest of the sessions with Professor Friedrichs.

5. *Quantum-Mechanical Hypothesis about the Brain, and the Idea of Replicating One*

What would follow if we entertain (for argument's sake) the hypothesis that the brain is so organized that for its working—and then for the behavior of the visible organism—transactions on the quantum-mechanical (subatomic) level can become relevant? Even before considering *how* they can do so, one surprising inference was pointed out by Friedrichs. I had spoken (above p. 221) of "a brain of the same physical constitution as the human, but without concomitant subjectivity," and then discarded the very hypothesis of a "merely physical brain," my point there being the nondissociability of a psychic complement from a living brain. But exception should be taken already to speaking of a "same" physical constitution as . . . with reference to the brain at all, which recalls the Cartesian hypothesis of exact "replication" treated earlier in this Appendix. For exact replication presupposes exact knowledge of all constituents of the system to be replicated. But if, *ex hypothesi,* such a system is really defined, in the causally meaningful sense, by quantum-mechanical state descriptions, then the exact, that is, complete, knowledge is on principle beyond reach and the very idea of an exact replication is inadmissible. It is—in the strict sense, and not just as a toll of excessive complexity—

impossible to have so exact a knowledge of a state of a human brain that we could predict all its future performances. Thus, also, one cannot design it, because in order to design it one must know it. But for what one *can* design there holds then the corollary truth that it will inevitably be a deterministic system. (This would not be basically altered by building in a random factor.) To quote Friedrichs: Our knowledge of the physical state of the human brain *can* only go so far that, in imitation of it, we construct a robot. In sum, if "state of the brain" means indeed "quantum mechanical state," then the famous Cartesian thought test of a perfect physical replication (by none less than the Creator, the perfect mechanic himself!) breaks down on this object. Thus, even without considering the role of the "subjective" factor (my original argument), robot would in terms of physical function alone, that is, in "overt behavior," always remain robot.

6. *Indeterminacy, Trigger Chain, and Macrobehavior ("Schrödinger's Cat")*

But how can "decisions" on the quantum-mechanical level determine macrobehavior? My model employed the trigger principle, and this was somewhat further elaborated by Friedrichs. First, the principle can be applied repeatedly and serially: triggering of trigger of trigger . . . , beginning with an arbitrarily small amount of energy. This I had in mind myself when speaking of the organism as an "amplifier." (Already Whitehead had done so.) But as long as classical physics is given unqualified validity, the trigger series will not lead back to something new: even its near-zero origin will still be subject to the general deterministic laws. However, "Schrödinger's cat" shows that with the trigger principle we can easily pass into regions where quantum physics takes over.

What is known as the problem of "Schrödinger's cat" was devised by him to illustrate the difficult problem of the role of the observer in quantum mechanics.[15] We shall use it here for a slightly different purpose, as it also illustrates, in a down-to-earth example, the difference of "causality" in classical and in quantum mechanics and at the same time shows a way in which the latter can be made to intrude into the domain of the former. The situation imagined by Schrödinger is the following. In a box, there is a cat, a vial of prussic acid, a sample of radioactive material, and a trigger mechanism that will break the vial (and thereby kill the cat) if, say, an alpha particle emitted by the radioactive substance hits a certain disk at the starting end of the trigger chain. If we know something of the state of the radiating material at a given moment, we may be able to compute with great accuracy the chances of the cat still being alive after one hour. Let us say that the probability is $1 \div 3$. What then can we say of the state of the cat after one hour, on the basis of our initial knowledge of the system at t_0 (i.e., ruling out our looking at t_1 through a window in

the box)? Only that the cat "is" one-third dead and two-thirds alive.[16] In classical mechanics we would have been able to predict the exact moment of the cat's death. But in quantum mechanics we cannot predict *when* "Schrödinger's cat" will be killed, nor later reconstruct when it was killed.

Now, of course, if there had been 1,000 boxes, with their initial conditions as nearly identical as they can be made, the pronouncement after one hour would not have been that absurd one (which it is with respect to one cat), but the entirely nonparadoxical one that two-thirds of the number are alive and one-third dead; and upon inspection, this would be found nearly true, the more nearly so the more boxes there are. Thus the quantum mechanically induced unpredictability of the single case vanishes again with increasing numbers. But in the question of "freedom and necessity" (or "mind-body" in general), not populations and averages matter,[17] but precisely the individual, just as with regard to the single cat it matters whether *it* is alive or dead, and *when* the latter contingency in the yes-no alternative takes place. The meaning of Schrödinger's thought experiment, as here used, was precisely to be a single-case experiment. Thus it illustrates what we are concerned with: "indeterminacy" carried over—by high subtlety of organization—from the micro- to the macroorder.[18] *If*, then, as the hypothesis has it, the human brain *is* such an organization, it may enjoy, for the macrodetermination of the body, that is, of our behavior (as well as for the internal determination of its nontransitive activities in mere thought) whatever latitude the quantum-mechanical indeterminacy of its base level offers it to play upon. This, to be sure, as Professor Friedrichs took care to stress, does not *explain* action of mind on matter or interaction between the two (there is, in his words, "no theory of that" in all this); but it does remove the standard objection that this whole notion is unacceptable to physical theory and the occurrence therefore to be denied. In other words, it disposes of the "incompatibility argument" in the psychophysical problem and thereby of the exclusionary dictate of materialism. The gain, even if lying in the negative, is philosophically significant: in quantum physics there is no flagrant contradiction between mechanics and the influence of consciousness.

This is how far the actual discourse, of which this is an expanded protocol, got by way of "results." If nothing else, they leave the feeling, or a persuasive conjecture, that it must be "here," in the *terra incognita* of the quantum-mechanical dimension, where the mysterious switch takes place—from mind to matter and from matter to mind (in its two-way interchange role an odd reincarnation of Descartes' "pineal gland" of unhappy memory). Moving beyond this *raisonnement* of possibility and the demarcation of its locus to an "explanation" or, rather, representation of the transaction itself would need a theoretical model whose terms are not borrowed from one side or the other—a *tertium quid*, neutral to the

distinction of matter and mind, prejudging neither one in the image of the other, but able to account for a transmutation, conversion—or whatever be the dynamical mode of transition—among the two. No such model is at present in sight. My all too crude metaphor of the osmotic "wall" could obviously not appeal to Friedrichs (nor does it much to myself). Something better is surely attainable, but a real theoretical solution may be destined to elude us.[19] It certainly did on this occasion. On this note of partial success and partial defeat, the discussion ended. As Friedrichs put it dramatically: "Having reached the point where I feel the solution must be looked for, I thought and thought—and finally threw up my hands."

Here, then du Bois-Reymond's *ignoramus* is presently lodged, perhaps even the *ignorabimus;* but better lodged than before, because freed of a spurious logical straightjacket. For myself (going back to what I said in n. 4), I add that even an *ignorabimus* in the terms of science need not halt in this matter the conceptual effort of speculative philosophy.

Appendix

1. "Brücke and I, we have bound ourselves by an oath . . .": so wrote du Bois-Reymond in a contemporary letter to Hallmann, quoted by W. W. Swoboda, "Ernst Brücke als Naturwissenschaftler," in Hans Brücke et al., *Ernst Wilhelm Brücke. Briefe an Emil du Bois-Reymond* (Publ. Archiv d. Univ. Graz, 8/1) (Graz, 1978), p. xxxiv. The writer speaks of Brücke and himself as "conspirators sworn to make prevail" the above truth and thereby transform physiology into an "exact science."

2. "Über die Grenzen des Naturerkennens," delivered at the forty-fifth Versammlung Deutscher Naturforscher und Ärzte in Leipzig, 1872; printed in E. du Bois-Reymond, *Reden* 1, pp. 441–473. The *ignorabimus* is there pronounced on *two* questions of natural knowledge: the psychophysical problem, and the intrinsic essence of matter and energy (as distinct from the laws of their action). Du Bois-Reymond thinks it possible that the two limits of our knowledge are at bottom the same, i.e., that if we comprehended the essence of matter and energy we would also understand how the substance underlying them will under certain conditions sense, desire, and think. "But [he says] it lies in the nature of things that on this point too we cannot obtain clarity, and all further talk about it remains idle" (p. 462). Thus, the last word of the essay, intellectually and typographically, is "*Ignorabimus.*" Against the protests of outraged scientific optimism that greeted his verdict at the time, we must still, a century later, attest to his philosophical insight in recognizing that these very questions are *transcendent to physics as such.* See n. 4 (below) for more on this question.

3. This view of the matter would, e.g., rule out telekinesis and other spiritistic macroeffects.

4. Trying for such a reformed ontology would, of course, be that kind of speculation of which du Bois-Rymond has said that no clarity (i.e., conclusive evidence) can be obtained on its subject and therefore all further talk, beyond indicating the mere possibility of a

common root for the divided record of things (and perhaps voicing a theoretical preference for its parsimony), must remain idle (see n. 2 above). But granted the first half of the contention, namely, that "knowledge" escapes us here, the second half concerning the idleness of the pursuit does not really follow, except by the terms of natural science and its defined criteria of verification. These do not exhaust the space of intelligibility and meaningful inquiry. "Speculative philosophy," says Whitehead, "is the endeavor to frame a coherent, logical, necessary system of general ideas in terms of which every element of our experience can be interpreted" (*Process and Reality,* part I, chap. 1, sec. 1). Reason must make that endeavor, even though foregoing in it the kind of verifiability which the positive sciences enjoy (in different degrees: history, e.g., far less than physics). After all, "epiphenomenalism" itself is a piece of "speculative philosophy," only a bad piece, because it fails at least two of Whitehead's tests: that of "coherence," and that of letting "every element of our experience" be interpreted in its terms. Usually, in the grand and forceful systems of speculation, one of these tests gives way to the other: most usually, in modern times, the second—fullness of interpretation of all phenomena—to the first, logical coherence, which since the seventeenth century has predominated in Western metaphysics with a certain surgical ruthlessness. At any rate, Whitehead's requirements, or such as his, do provide standards for judging a speculative scheme and save the enterprise from mere fancy. To interpret reality in the light of *all* the knowledge we have of it, transcending a mere summation thereof, is a need, a right, and a duty of reason, and is different from exploring its particular provinces. The "need" was acknowledged by none more movingly than by Kant, but he denied reason the right to follow its innate thirst because no *knowledge* to quench it lay on its path. Whitehead, instead, speaks of "interpreting," and this—not quite the same as "knowing"—is not only unavoidable (for we do it anyway and on whatever level, primitive or sophisticated, articulate or inarticulate), but it also remains meaningful in the face of nondefinitiveness and without the blessing of ascertained truth. To return to our special subject: the psychophysical problem is one of the pressures with which multiform reality nudges reason beyond the safety of the severally uniform sciences into the quest for transcendent unity that can only be proposed and never proven; and after Descartes it happened to be the major such pressure to which the great metaphysical systems responded with their different solutions, all of them tinged with the characteristic violence of thought to which I have referred. More heedful to the shadings of experience, yet bolder by the radical conceptual reframing of ontology to do them justice, is Whitehead's grand attempt in our century. Its persuasiveness, besides the power of its internal coherence, depends on the measure to which it passes the external test set by himself—the oldest in metaphysics: "saving the phenomena," i.e., whether all of them can be interpreted without loss of character in terms of the system. It is no belittling of Whitehead's achievement, but a call to go on, to say that *not every* significant element of our experience is so "saved" in his conceptual scheme. But it is an inspiring instance, the only one so far, of what I have in mind when, over against my feigned "speculative model" framed in conventional terms, I surmise a truer one "framed in ontological terms which would alter our very speech of 'matter' and 'mind.' "

5. With reference to this aspect, one well may think of Niels Bohr's utterance of 1952 as reported by Werner Heisenberg: "If one is not at first shocked by quantum theory, one cannot possibly have understood it." Bohr missed that "shock" in the response of a philosophers' meeting in Copenhagen—mostly of the positivistic persuasion—to a talk by him (see Werner Heisenberg, *Der Teil und das Ganze. Gespräche im Umkreis der Atomphysik* [Munich: R. Piper & Co., 1969], p. 280).

6. Ideally, the state of the system at a given moment consists of all possible "observables," i.e., of whatever one can measure: position, momentum ("velocity"), spin, etc. All observables together, known simultaneously for time t_0, constitute what we may call the "Laplace state," which represents the ideal of classical physics. With that knowledge, held

possible in principle, all future and past states are also known, namely, determined. Just this simultaneous knowledge is held on principle to be impossible by quantum theory, notwithstanding the knowability of each constituent by itself. The very actualizing of the knowledge in one direction forfeits that in another (see continuation in text).

7. "Ultraphysical" renders the German *ultraphysikalisch* = beyond the grasp of physics, an epistemic term, not (of course) *ultraphysisch* = beyond corporeal nature—an ontological term which would place the "reality" in question in a different realm of being, e.g., the mental (in which case it would be eminently knowable!). The English "physical" collapses these two different meanings into one equivocal term. The nearest analogy in the philosophical vocabulary to the intended meaning may be Kant's "noumenal," referring to "the thing in itself," whose formal *concept* belongs to the "intelligibles" (= formed by the intellect alone), but whose content we cannot obtain. "Ultra-" was chosen in deliberate preference to "trans-" because of the latter's strong connotation of a transcendent, qualitatively different *kind* of reality (like meta-physical), whereas "ultra" can also mean more of the same, exceeding it in its own kind, as in "ultraradical," "ultraconservative." (This semantic clarification was provoked by the doubts of one very attentive reader of the manuscript, Professor Adolf Lowe.)

8. With the remainder of this section I depart from the digest of the talks, where this retrospective theme did not come up.

9. See, e.g., Niels Bohr, *Atomic Theory and the Description of Nature* (Cambridge, 1961), pp. 24 and 100 ff.

10. For the particle-wave alternative the conflation of epistemic with ontologic meaning, audacious as it is, is at least arguable, as the "objects" in question are *entia rationis* (theoretical constructs) to begin with. E.g., the wave aspect of the quantum-mechanical event, though mathematically isomorphous with the description of a concrete, physical wave, denotes so highly abstract an "object" as a probability wave, whose reality status and existential independence from the conceptualizing observer are indeed debatable. Nothing like it applies to such concrete, content-saturated data of our primary experience as body and mind.

11. This goes for both sides, though it is more obvious for the mental side by which we exemplified it: a sense perception is *of* a physical object and also (we assume) caused *by* that object. The physical side seems better isolable: one *can* describe the physics of the eye without any reference to seeing, and so with every part and even the whole of the physical organism. But I doubt whether it makes sense to do so for long (e.g., to treat of brain processes and not mention their mental implication), and even whether it is possible to speak of the quality-stripped entities of physics in general without in the negation *implying* the perceptual qualities from which they were abstracted. Also we must say, after all, that the pigment on the canvas causes the color perception in us to which it is correlated: this *acting on* the mind is as surely a statement about the *physical* thing as the being affected by it is a statement about the mind. In short, the two "sets," whether the one or the other is thematized, keep crossing over (or "interact"), and this is just the crux of the matter.

12. Even simpler, in this formalistic vein, is the objection that in the quantum-mechanical case we begin with data of the same kind (space-time measurements) and end up with a duality of representation of our own devising to account for them, whereas in the psychophysical case we begin with a duality of cardinally different data, not of our making at all, and try for a theoretical unification of them. If in such a unification they are found to be "complementary" in some sense, then "complementarity" itself becomes the unitary representation for a dual phenomenon. Thus, the direction of the logical operation in the two cases is opposite, the one yielding a divergent model, the other (hopefully) a convergent one. These purely formal objections, by the way, especially that of the "crossing lines," fall on *all* the extramural uses of the complementarity principle I know of (e.g., in the social

sciences): they all are forced to violate the (at least) *semantically exclusionary* character the duality has in the original model and come to grief already on this count alone.

13. It is equally meaningful to ask *what* of our behavior, even of the mental state in back of it, is conditioned or circumscribed or prescribed by physical necessity, and what we truly initiate—i.e., to *apportion* the relative *shares* of the two sides in a given instance: again something wholly inapplicable to complementarity in the genuine sense.

14. For Spinoza, I have shown this in an article, "Parallelism and Complementarity: The Psychophysical Problem in Spinoza and in the Succession of Niels Bohr," in *The Philosophy of Baruch Spinoza*, ed. R. Kennington (Studies in Philosophy and the History of Philosophy, vol. 7) (Washington, D.C.: Catholic University of America Press, 1980), pp. 121–130 (title there mutilated by typesetter's error). The discussion of complementarity there is mostly identical with the present one.

15. E. Schrödinger, *Naturwissenschaften* 23 (1935): 807.

16. To the objection that no statistician would dream of making such a statement about person x on the basis of a life expectancy table for the population (but would say that no pronouncement at all is possible on the single case), the answer is that the two cases are not analogous. In normal statistics, the knowledge is about a population, and so, of course, are the predictions based on it. None on individuals are to be expected; but in "Schrödinger's cat," the initial knowledge is precisely of the state of the individual system (in the optimal case: as exhaustive a knowledge of "compatible" quantities defining it as on principle can be had together), and so predictions on the future state of *that* very system *are* to be expected and indeed are provided by the Schrödinger function. The proper analogy, therefore, is not between the cat and person x, but between the cat and the population to which x belongs—and there the prediction that two-thirds of "it" (namely, of its unspecified present members) will be alive and one-third dead at t_1 is perfectly meaningful and, maybe, true. But for the indivisible cat it is "true" only as a teaser. For both, of course, the simple and nonproblematical statement would do that the odds for the individual to be alive are two-thirds, and one might leave it at that, but then would blanket a profound difference. In the ordinary statistical case the statement merely expresses (e.g., to the insurance company) the individual's membership in the sample from which the probability ratio had been averaged and is in no sense a *causal* statement: no dynamic analysis of the "state" of either whole or part is involved. In the Schrödinger case the statement does express precisely the internal—and intrinsically probabilistic—"causality" of the analyzed individual system-state itself: a hardly comparable situation.

It is this difference, and the unorthodox nature of prediction under quantum-mechanical, probabilistic conditions as opposed to those of classical mechanics (under which population statistics are fully comprised) that the paradoxical—admittedly facetious—expression chosen by my interlocutor was to convey. It would be lost by submerging it in the classical relation of large numbers versus individual instances, which does yield precise (though not causally derived) predictions—for the large numbers. (This again is in reply to objections raised by Adolf Lowe.)

17. In population statistics we can allow a large measure of determinism concerning rates of mortality, births, crime, etc., without determining thereby any single cases, e.g., whether and when *I* die, procreate, commit a crime. The single cases are taken to be *diversely determined* by clusters of individual causes of their own—known or unknown, but anyway ignored—which average out in the large enough population sample. Note again the difference from the seemingly identical outcome in the Schrödinger example. There, to obtain that outcome, we had to stipulate for the parts initial conditions as nearly identical as they can be made. No such condition is imposed on ordinary statistics. There, on the contrary, the definite ratio results for the whole from the confluence of indefinitely *different* initial conditions (and consequent causal histories) in the parts. No such differences are averaged out in the multiplication of Schrödinger's cat, rather are identities converted from probabilities

in the parts toward certainty in the whole. The ratio coming out at the end was there in each of the parts at the beginning, and what is averaged out—toward determinacy—by the large number is merely the many, equal-valued indeterminacies, not many different determinations toward a mean value. Or, the mean value is given by the probabilistic equation of the single case. (In an analogous "classical" case, e.g., if I know of a population of insects that they all are ruled by the same biological clock, I do not, when I know its setting for one specimen, need statistics to tell me that the whole population—number irrelevant—will die at the onset of winter. But then, the individual biological clock is not probabilistic.)

18. This is technically not only feasible but quite familiar: any Geiger counter registering single radiation particles is an instance of it.

19. One attempt known to me from the camp of quantum physics, which avoids the temptation of complementarity and squarely tackles interaction, is L. Bass, "A Quantum Mechanical Mind-Body Interaction," *Foundations of Physics* 5/1 (1975): 159–172. It takes its cue from the introduction of the *consciousness* of an observer as an essential part of one version of the postulates of quantum mechanics (von Neumann), according to which one can say, "The impression which one gains at an interaction, called also the result of an observation, modifies the wave function of the system" (E. P. Wigner). In "a short step further," Bass imagines the inanimate microsystem thus modifiable to be placed in a strategic element of the central nervous system of the observer himself, so "that a suitable wave function pertains to a nerve cell which undergoes excitation when the wave function is modified by an event in consciousness." In this way, the connection between mind (= observer) and microscopic systems, posited in the general theory, is "brought to bear on the possible connection between the mind and the body of the observer" (p. 160). A formal model of such an element in the central nervous system is then constructed by Bass. Crucial for its working is, of course, that the element in question is "observed." Since, evidently, it is not so by the consciousness of the person, Bass resorts to the introduction of "subbrains," invoking C. S. Sherrington's view of the mind as "a collection of quasi-independent perceptual minds integrated in large measure by temporal concurrence of experience" (p. 170). Any such "subbrain," recording the relevant datum as a "quasi-independent mind," may then play the role of the "intermediate observer," and the "integrated mind" of the subject the role of the "ultimate observer" who records the state of the neural net *in toto*. Among quantum physicists this "ultimate observer" is known under the name "Wigner's friend," since it is borrowed from a thought experiment by E. P. Wigner.

The internal mathematical soundness of the model construction is beyond my judgment. I also waive the objection that we have no evidence of the existence of "subbrains"; and that anyway what we need would be "subminds"—by no means the same thing; and that to speak of mind parts, as one can perfectly well speak of brain parts, is questionable in itself, and the inference to a part-mind from a local brain part (even if this *were* a part-brain) is questionable once more: I waive this type of objection, because widest largesse shall be allowed to speculation at this stage. Not to be waived is the objection that the basic premise of the model construct confounds or conflates, under the blanket term of "consciousness," two different things: the state, or event, of subjective awareness and the objective act, or event, by which it was obtained, namely, measurement, which is a *physical* intervention in the state of the object. Here, Wigner's language—in the above quotation—is ambiguous. "The impression which one gains at an interaction . . . ": Is it the gaining or the having of the impression that modifies the wave function of the object? Surely it is the former, as a case of true physical interaction. Otherwise, the requirement of simultaneity is waived and we could even modify, as late observers, the wave functions of systems in the remote past, inasmuch as we gain knowledge of them via the intervening history of the universe. Nobody can mean that our ideas of the past can change the past. But so long as we stick to simultaneity of observer and observed and then place the modifying of the object in the *extramental* activity of *gaining* information, where by quantum-mechanical rights it

belongs, we have not left the strictly physical sphere and still do not know how the mental datum once there, the information gained and held in consciousness, *then* in turn can become the fountainhead of a causal traffic with the object that is to be swayed *because* of (in answer to) "information received." Not what in the act of perceiving we may unknowingly have done to the object, but what, possessed of the percept, we may now or later knowingly do to the object—acting on it with our bodies and thus in the first place *making* our bodies so act: that is the causal hub of the mind-body problem. The initial puzzle still stares us in the face.