

THE METAPHYSICS OF MENTAL CAUSATION: EXPLANATORY EXCLUSION REDUX

Robert C. Koons
Department of Philosophy
University of Texas, Austin

The problem of mental causation is the Achilles heel of physicalism. In this paper, I identify and defend the metaphysical grounds for a cogent challenge to physicalistic mental causation, by arguing that Humean accounts of causal explanation (required by physicalistic mental causation) are inconsistent with the intrinsicity of causal-explanatory connections, and that the completeness of the microphysical domain (required by a coherent physicalism) is inconsistent with the exact truth of any kind of functionalist theory of the mind.

THE METAPHYSICS OF MENTAL CAUSATION: EXPLANATORY EXCLUSION REDUX

Robert C. Koons
Department of Philosophy
University of Texas, Austin

1. Introduction

On the question of causation, philosophy of mind has, until very recently, lagged behind the pace of metaphysical developments. In particular, the conception of causation incorporated into Functionalism and non-reductive physicalism more generally was a thoroughly Humean one.¹ Now that anti-Humean theories of causation are available, conventional wisdom about the mind/body problem, and the related problem of free agency, is in is long overdue for a thorough re-examination.

Jaegwon Kim was one of the first philosophers to recognize that the problem of mental causation is the Achilles heel of physicalistic Functionalism. In fact, the situation is even worse than Kim imagines: even if mental properties were reducible to disjunctive or macroscopic physical properties, mental causation would fail. In this paper, I specify and defend the metaphysical grounds of a cogent challenge to physicalistic mental causation.

2. What's Wrong with Functionalism

In David K. Lewis's version of Functionalism,² mental states can be identified with certain logically complex, "higher-order" states, states definable in terms of existential quantification over "first-order" physical states. Lewis proposed "Ramseyfying" some theory Ψ of the causal roles of mental states (including their dispositions to be caused by

¹ By "Humean", I mean the 19th and 20th century traditions in English speaking tradition that adopted Hume's idea of "constant conjunction" as an analysis or metaphysical account of the nature of causation. I don't mean to imply that Hume himself was a "Humean" in this sense.

certain environmental conditions, their dispositions to produce other mental states, and their dispositions to cause behavior). The Ramsey technique involves replacing the mental-state expressions in Ψ with variables, in particular, with second-order variables $X_1 \dots X_n$ that range over a domain of possible physical states. For example, the predication of the k th mental-state expression in theory Ψ to an individual c would be transformed into the following:

$$\exists X_1 \dots X_n (\Psi(X_1 \dots X_n) \ \& \ X_k(c))$$

The relation between a mental-state so conceived and the physical state that “realizes” it is a straightforwardly logical one: a physical state $\phi_k(c)$ realizes the mental state corresponding to the k th expression in the theory Ψ just in case ϕ_k is one of a set of physical states ϕ_1 through ϕ_n that jointly satisfy the expression ‘ $\Psi(X_1 \dots X_n)$ ’.

The teleological account of the mind developed in recent years by Dretske, Stampe, Millikan³ and others can be seen as a version of Lewis/Ramsey Functionalism. The etiological account of proper function that these have offered, building on work in the 70’s by Larry Wright, provides some details about the form of the mental-state theory Ψ . In particular, each mental-state is identified with a particular proper function, and a proper function of a state is defined in terms of the causal history of that state. More specifically, a physical state of form ϕ is said to have α as its proper function just in case (i) there is a causal-explanatory connection between states of type ϕ and states of type α , and (ii) this ϕ -to- α connection played a role in the actual causal history of this particular ϕ -state.

² “Psychophysical and Theoretical Identifications,” *Australasian Journal of Philosophy* 50 (1972):249-258.

³ D. Stampe, “Towards a causal theory of linguistic representation,” in P. French, T. Uehling and H. Wettstein, eds., *Midwest Studies in Philosophy*, vol. 2, *Studies in Semantics* (University of Minnesota, Minneapolis, 1977); Fred I. Dretske, *Explaining Behavior: Reasons in a World of Causes* (MIT Press, Cambridge, Mass., 1988), Ruth G. Millikan, *Language, Thought and Other Biological Categories* (MIT Press, Cambridge, Mass., 1984);

Lewis/Ramsey Functionalism depends on the kind of view of mental causation expounded by Donald Davidson in “Mental Events”.⁴ Davidson assumes that the physical domain is causally complete, and, therefore, that mental states can be efficacious only if they are in fact identical to physical states of a certain kind. When a mental state causes something, it always does so qua physical state. Similarly, according to Lewis Functionalism, it is always physical states that enter into causal relations. A mental state is simply a higher-order physical state, the state of there being some actual physical state with such-and-such causal properties. Mental states themselves do not enter into causal explanations.

As Jaegwon Kim has pointed out in a series of articles and books, this explanatory exclusion of the mental by the physical is difficult to square with our prephilosophical convictions concerning the efficacy of the mental.⁵

3. The Semi-Humean Response

However, as Barry Loewer has recently pointed out,⁶ this problem of explanatory exclusion goes away if we take a Humean, or even semi-Humean approach to causal explanation. If causal explanation is to be understood as a modal or probabilistic relationship between two categories (and it doesn’t matter if we take probabilities, as a strict Humean would, to be supervenient on actual frequencies, or if we have a more robust notion of objective propensities), then there is no reason why mental states can’t figure in genuine causal explanations, even for a physicalist.

Take, for example, a very simple example of a Functional state, that of a disposition, like solubility. Being soluble can be identified with the higher order state of being in some

⁴ “Mental Events”, in *Essays on Actions and Events* (Clarendon Press, Oxford, 1980), pp. 207-27.

⁵ *Supervenience and Mind* (Cambridge University Press, Cambridge, U. K., 1993); *Mind in a Physical World* (MIT Press, Cambridge, Mass., 1998).

⁶ “Review of *Mind in a Physical World* by Jaegwon Kim,” forthcoming in *Philosophy and Phenomenological Research*.

physical state that would very probably cause, in the presence of water, the formation of a solution. For a Humean, the solubility of a substance could be part of a genuine causal explanation, since the right kind of probabilistic relationship holds between the state of being soluble and in the presence of water, on the one hand, and the subsequent formation of a solution, on the other. In *Realism Regained*,⁷ I gave just such a semi-Humean account of causal explanation, concluding that genuine mental efficacy does not depend on any lack of causal completeness at the physical level.

On a Davidsonian account of mental events, mental events have two aspects: physical and mental. For Functionalists, these two aspects correspond to two different predications applicable to the event: a predication of a first-order physical state, and a predication of higher-order physical state, that is, the state of being in some first-order physical state with certain causal characteristics. For Humeans and semi-Humeans, both aspects can figure in genuine causal explanations, since the extension of the higher-order state can stand in the right modal or probabilistic relations to other extensions.

Even if we adopt a resolutely anti-Humean account of the causal relation between concrete events, we can still adopt a semi-Humean account of the causal-explanatory relation between the more abstract event-aspects (as, again, I did in *Realism Regained*). We can believe in singular causation between events, arguing that the presence or absence of a causal connection between two events is not determined by the non-causal properties of the two events, and still insist that the presence or absence of a causal-explanatory relation between the event-aspects of two events is wholly determined by general, probabilistic relations between the two general properties definitive of the two event-aspects. For example, we might insist that the existence of a causal connection between two mental events is a primitive and irreducible matter, while also asserting that the fact that the mental aspect of the first event causally explains the physical aspect of the second is reducible to two sets of facts: (i) the fact that a causal connection does in fact hold between the two events, and (ii) facts about the objective probability of the

⁷ *Realism Regained: An Exact Theory of Causation, Teleology and the Mind* (Oxford University Press, Oxford, 2000).

existence of an effect with the corresponding physical property, given the existence of an event with the mental property.

By a “semi-Humean account”, I mean an account of causal explanation according to which the combination of two things is *sufficient* for the existence of a genuine causal explanation between the property-instances of two events: (1) the existence of a causal linkage between the two events (as concrete particulars), and (2) a modal or probabilistic relationship between the two property-types. This does not make just any account of probabilistic causation semi-Humean. For example, a realist account (along the lines of Dretske, Armstrong and Tooley), which grounds probabilistic causal explanations in a stochastic connection between two universals, is not even *semi-Humean*, since the existence of a probabilistic relationship is not *sufficient* (on these accounts) for the existence of a causal explanation: the probabilistic relationship must be *directly* grounded in a real connection between the two properties. For example, if there is a stochastic tie between A and B, and between A and C, then there will be probabilistic relationship between A and and the disjunction of B and C (the presence of A will raise the probability of the disjunction by raising that of each disjunct), but occurrences of A will not count as causal explanations of occurrences of (B or C), because there is no direct causal link between A and (B or C) (considered as universals).⁸

4. Four Problems with the Semi-Humean Response

What, from a metaphysical point of view, is wrong with Humean and Semi-Humean conceptions of causation? Recent work in metaphysics and philosophy of language have given causal explanation an increasingly central role: we have, as the leading theory, or at least one of the principal contenders, causal theories of reference and representation, of perception and knowledge, of diachronic identity (both personal and material), and of

⁸ According to most anti-Humeans, such disjunctive properties as (B or C) do not even exist: i.e., there is no corresponding universal. Nonetheless, even in cases where two universals do exist and there is probabilistic relationship between their extensions, this relationship cannot ground a causal explanation unless the relationship is grounded directly (and not derivatively) in a fundamental stochastic tie between the two universals.

space and time. This increasing centrality of causation suggests that causation must be admitted as one of the fundamental building blocks of reality, as “the cement of the universe,” as J. L. Mackie put it. Such universal cement must be an intrinsic feature of the connections it establishes. Slogan: *Fundamental relations must be intrinsic relations*. We have, then, good reason to accept the following principle of the Intrinsicity of Causation:

(IC) Causal-explanatory connections are intrinsic features of pairs of events.⁹

By saying that the causal-explanatory connection is intrinsic to the cause/effect pair, I do not mean to beg the question of whether causation is an internal/logical or external/real relation: that is, whether causal relations supervene on the individual natures of the two events or represents something over and above this. What I do mean to assert is that the causal-explanatory connection between aspects of a cause and of an effect does not consist in anything extrinsic to the particular pair so related. Any causal connection between a cause and its effect is wholly intrinsic to the local situation consisting of the two events and their relations to one another.¹⁰

But what do I mean by “intrinsic”? David Lewis proposed that we define intrinsicity in terms of exact duplicates: a property is intrinsic if and only if it is shared by all exact duplicates. However, this clearly gets the order of explanation backward: two things are duplicates because they share all their intrinsic properties, so it cannot be the case that two properties are intrinsic because they are shared by all duplicates. Instead, we should understand intrinsicity in terms of mereology, the part-whole relation. Something is intrinsic to a situation just in case it is literally part of that situation, or if it is connected to the situation by a so-called internal or logical relation, like that between something and

⁹ I am speaking here of direct or immediate explanatory connections. If a pair of events is connected by a chain of events, then the causal-explanatory connection will be intrinsic to the *chain* and not just to the pair of events constituting its endpoints.

¹⁰ All “internal” (or logical) relations are intrinsic, but not vice versa. A real/external relation can be intrinsic, so long as it involves nothing beyond the pair-in-relation (at least, nothing except things that are “internally” related to the pair).

a Platonic universal that it participates in. In other words, there are only three things intrinsic to a given situation: things that are literally parts of the situation, Platonic universals that are exemplified in the situation, or intrinsic relations between those same Platonic universals. (If we reject a Platonistic conception of universals, then we can identify simply intrinsicity with parthood.)¹¹

1. Problem 1: the intrinsicity of event-aspects

Since causal explanation is a relation between the aspects of two events, the causal-explanatory connection can be intrinsic to those two events only if the event-aspects are themselves intrinsic. Thus, the principle of the Intrinsicity of Causation has the following corollary, the Intrinsicity of Event-Aspects:

(IAE) Event aspects are intrinsic to their events.

However, on the semi-Humean account, the relation between an event and one of its aspects is instead an *extrinsic* one. In other words, for the semi-Humean, there is no sense in which the event-aspect can be thought of as a genuine part of its event.¹² For the semi-Humean, the existence of an event-aspect is reducible to some extrinsic fact about the event: either, its belonging to the extension of a particular linguistic expression, or its belonging to a particular kind of class (such as a “natural class”, in Lewis’s later work), or its bearing an extrinsic similarity relation to certain exemplars.

¹¹ If Platonic universals can themselves instantiate other Platonic universals, then the definition of intrinsicity will have to be recursive, with being-part-of as the base case. In other words, a relation *is intrinsic* to a pair if it is part of the local situation consisting of the pair, or is part of a situation relating two universals instantiated by the pair, or part of a situation relating two universals, each of which is instantiated by a universal instantiated by one of the pair, or...

¹² For an account of the ontology of event-aspects and its impact on the logic of causal explanation, see my “The Logic of Causal Explanation: An Axiomatization,” forthcoming in *Studia Logica*.

However, genuine causal explanation ought to be a relation wholly intrinsic to the two events. If aspect A of event C causally explains some effect E, then A must be a genuine constituent of the event C. The reality of A's distinctive causal contribution must be grounded in a distinct component of C, a requirement that semi-Humean accounts cannot meet.

2. Problem 2: The intrinsicity of nomological connections.

For the semi-Humean, causal explanations are grounded in a relation between two general properties or universals. This can be understood in one of two ways: a fully Humean approach, in which the relation is defined in terms of actual frequencies of different kinds of events, or a semi-Humean approach, in which the relation is defined in terms of necessitation or objective chance, where facts about modality and chance are not assumed to supervene upon the distribution of occurrent qualities in the actual world.

The semi-Humean approach introduces, in addition to the actual world, an array of merely possible worlds, distributed across logical or modal space. Objective chance is determined, not by relative frequencies in the actual world only, but by relative frequencies across the whole of logical space, or, at least, across some relatively large neighborhood of the actual world. Nonetheless, in the end, nomological connections (that is, connections involving objective chance or physical necessitation) consist merely in regularities. This is true for the semi-Humean as well as for the Humean. For the semi-Humean, the relevant regularity extends beyond the bounds of the actual world, but it is still merely a regular association of occurrent qualities. For both the Humean and the anti-Humean, causal laws are merely *accidental* generalizations of a certain kind: for the Humean, accidental generalizations that form part of a simple axiomatization of the actual world's history; for the semi-Humean, accidental generalizations that hold, not only throughout the actual world, but throughout a range of nearby possible worlds.

Thus, for both the Humean and the semi-Humean, nomological connections between causes and effects are radically extrinsic to those events. Since, for both the Humean and

the semi-Humean, the presence of a nomological relation is, at least, an essential component of the causal connection, on the Humean and semi-Humean approaches, the existence of a causal explanation is almost entirely extrinsic to the pair of events involved. Whether one aspect of one event causally explains another depends on many remote facts about the history of the world, or even more remote facts about other possible worlds. Hence, the intrinsicity of the causal connection is inconsistent with the semi-Humean approach.

The intrinsicity of causation entails the existence of a causal-nexus fact that relates the aspect of one event to the aspect of another. This could happen either at the level of tropes (abstract particulars or particularized properties) or at the level of universals. On a trope theory, each case of causal explanation would involve some intrinsic relation between two tropes, presumably by virtue of the presence of a third, relational, causal trope connecting the two. On the universal theory, we would have something like the Dretske/Armstrong/ Tooley account of causal laws, with causal explanation consisting in a second-order universal connecting the two universals corresponding to the two event aspects. In either case, the relation of causal explanation is intrinsic to each cause/effect event-pair. On the universal account, the relation of instantiation between the cause/effect pair and the causal law must be a logical or “internal” relation, with the universal being fully present in each particular instance.

In addition, if we adopt a Dretske/Armstrong/Tooley account, we must distinguish between two kinds of modal and probabilistic connections: basic connections and derived connections. For example, if we have a basic modal connection by which F necessitates G, and another by which G necessitates H, then we have a derived connection between F and H. Instances of F will have to be instances of H, but not because of any basic connection between F and H, but only because of the two basic connections linking F to G and G to H. Similarly, if there is a basic necessitation connection between A and C, and another between B and C, then there will be a derived necessitation relation between the disjunctive property (A or B) and C. However, there will be no basic connection between the disjunction (A or B) and C. If an instance of A gives rise to an instance of C,

it is the property A, and not the disjunctive property (A or B), that bears the weight of causal explanation. Causal explanation always coincides with basic connections between universals; never with merely derived connections.

Traditional modal logic, with its domain of possible worlds, can be taken as a tool for exploring the logical constraints on possible sets of basic and derived modal facts. However, traditional modal logic cannot by itself capture the distinction between basic and derived modal facts. Hence, traditional modal logic is inadequate to the task of explicating causation and causal explanation, since causal explanation exists only when supported by a basic modal or probabilistic connection.

3. Problem 3: No ontological free lunch

There is, within the semi-Humean account, an ineliminable tension between the claim that the introduction of higher-order states is an “ontological free lunch” (in Armstrong’s phrases), i.e., nothing over and above the first-order facts, and the claim that these same states can be load-bearing supports of causal explanations. If we take the deflationary, “nothing-over-and-above” line, then there is simply nothing there in the particular event to play a role in a distinctive causal explanation. If, to the contrary, we take seriously the positing of the higher-order state as a first-class citizen of our ontology, then we must locate it somewhere in the causal network. Such states must be either epiphenomenal (which explicitly sacrifices mental causation) or shoe-horned somehow into the causal history of physical events (which explicitly sacrifices physicalism). If the physical domain is conceived of as physically closed, however, it will have to be thought of as impermeable to such higher-order intrusions.

Consequently, not only Functionalism, but any version of supervenience theory committed to the causal closure of the physical will be inconsistent with mental causation. As Daniel Bonevac has argued, the supervenience of the mental on the physical amounts simply to a reduction of the mental to the physical “in the mind of

God”.¹³ Each mental property would consist merely in an infinitely long disjunction of physical properties, but, as I have argued, such disjunctive properties cannot, without violating the causal closure of the physical, bear any causal-explanatory weight.

D. Problem 4: The irrelevancy of remote facts

When one factum A causally explains the obtaining of a second B, A is responsible for producing B, for bringing B into existence. Facts that are remote from A and B, by which I mean facts that are wholly causally prior to A, causally posterior to B, or causally unrelated to B, cannot be involved in A’s production of B. Consequently, A’s causally explaining B cannot involve such remote facts. If it did, then they would be, contrary to the hypothesis of their remoteness, causally prior in an immediate way to B itself. However, according to Humean and semi-Humean accounts of causal explanation, A’s causally explaining B does consist, in part, in a very large number of remote facts about the regular association of A-type events with B-type events (either in remote parts of the actual world, or in nearby possible worlds). Hence, Humean and semi-Humean accounts are inconsistent with the irrelevancy of such remote facts to causal connections.

In fact, Humean and semi-Humean accounts play absolute havoc with the causal structure of the world. A great many events that are causally posterior to B, both in this world and in other possible worlds in which B occurs, are relevant, according to these accounts, to the production of B by A. Since this relevancy implies that these events are also causally prior to B, these Humean accounts entail the existence of innumerable causal loops.

Of course, the Humean would insist that, according to his proposed definition of “causal relevance”, these remote facts are **not** *causally relevant* to B. However, they are, according to the Humean account, metaphysically relevant to A’s production of B (the fact of A’s production of B consists in the obtaining of these facts): hence they are prior to B in the order of metaphysical explanation. In fact, B is prior to itself in this order.

¹³ “Reduction in the Mind of God”, in *Supervenience: New Essays*, edited by Elias E. Savellos (Cambridge, Needham Heights, 1995).

Since A's power to produce of B is, according to the Humean, dependent on a correlation (either in this world alone or in this world and nearby ones) between A-like events and B-like events, whether A is followed in the actual world by B is relevant to the question of whether A has the power to cause B. However, this is an incoherent position: it does violence to our concept of *causation* or *production* to suppose that *A's power to produce B* is dependent on the actual occurrence of B. In reality, if A produces B, then the occurrence of B in the actual world is wholly dependent on A's power of producing it: hence, the fact that A has the power to produce B cannot depend on B's actual occurrence, contrary to the Humean position.

5. A Resolutely Non-Humean Alternative

A popular idea in recent philosophy, the introduction of so-called 'truth-makers', can be enlisted in the construction of a non-Humean alternative. These truth-makers (called 'facta' by D. H. Mellor) are concrete parts of the world that are responsible for grounding the truth-values of statements and propositions. They can be conceived of as either situations or states of affairs (something like the atomic facts of the logical atomism of Russell and Moore) or as tropes (abstract particulars, scholastic individual accidents). For my purposes here, further specification of these facta/tropes is not needed.

The relata of causal explanation are event-aspects. We can identify events with certain aggregates of facta or tropes, and event-aspects with certain parts of these aggregates. Roughly, an event is an aggregate of tropes with a common causal history, and an event-aspect is a part of such an event possessing a distinctive causal power. According to this approach, the difference between two causal explanations appealing to different aspects of the same event has a real, ontological foundation. A distinctive mode of causal explanation is borne by a distinct part of the relevant event.

If, on this non-Humean view, there are non-physical aspects of events that genuinely enter into causal explanations of physical events, then the physical domain cannot be causally complete. This means that physicalism is inconsistent, not only with mental

causation, but with causation associated with any of the special sciences (i.e., with anything except micro-physics).

6. Yablo's Way Out: Biting the Overdetermination Bullet

Anti-Humean physicalists can avoid the exclusion problem by going to the opposite extreme from ontological minimalism. For example, Yablo¹⁴ and Shoemaker¹⁵ have recently advocated a plenitudinous theory of events, in which distinct tropes or facta are posited for every true disjunction or generalization, however arbitrarily constructed. On this approach, there are plenty of concrete events to bear real causal-explanatory relations, but the Yablo-Shoemaker account entails the existence of the very kind of massive overdetermination that Kim has taken such pains to avoid. Each highly abstract event (disjunctive or existentially generalized) is at least doubly caused on the Yablo-Shoemaker account: once on a plane of equally abstract causes, and again by being a logical consequence of more concrete events, which are themselves causally explainable on a plane of equally concrete causes. A plenitudinous ontology of this kind leads to plenitudinous overdetermination.

Yablo avoids this result by defining causation in such a way that the causes of the physical realization of a higher-order state does not count as a "cause" of the supervening, higher-order state. This is not because the purely physical cause has too little information to explain (as well as can be explained) the resulting higher-order state: it is rather that the physical cause has too much, extraneous information that it is disqualified as a "cause" of the higher order state.

Although Yablo's move avoids at a verbal level the conclusion that the higher-order state has multiple simultaneous "causes", it does nothing to remove the very real overdetermination of the higher-order state. The purely physical level still provides us

¹⁴ Stephen Yablo, "Mental Causation," *Philosophical Review* 101 (1992):215-280.

with a maximally good explanation of the subsequent realization of the higher-order state: its only “defect” is that it explains *more* than it must if it is to qualify as a Yablo-cause of the state.

I may be misinterpreting Yablo: perhaps he does not intend that every disjunction count as a “determinable”, of which any disjunct within it counts as a relevant “determinate”. The crucial question is this: what, if anything, is the causal/explanatory relation between a determinable and a corresponding determinate? If there is no causal relation, or if it is the determinate that causally explains the determinable, then the determinable (the mental state) does no real causal-explanatory work. It is the determinate that fully explains whatever is explained by either, in that case, rendering the determinable causally superfluous. In order to count the determinable as causally explanatory, we would then have to overpopulate the world with a vast number of redundant causal connections.

If, in contrast, Yablo were to suppose that, whenever both a determinable and a corresponding determinate actually occur, it is always the determinable that causally explains (at least in part) its determinate, that is, if the determinable (e.g., the mental state) is always causally prior to its determinate (the physical state on which it supervenes), then his account would be in direct conflict with the physicalist assumption that the physical is causally closed, and Yablo’s account would be immune to the causal exclusion problem that besets such physicalism

7. Another Problem with Functionalism

There is another problem with the wedding of Functionalism with physicalism. I will first illustrate this with the case of teleofunctionalism of the Dretske-Millikan variety. Teleofunctional accounts of proper functions assumes that gross, macroscopic properties can be causally explanatory. For example, the teleofunctionalist's explanation for why the

¹⁵ Sydney Shoemaker, “Realization and Mental Causation,” in *Physicalism and its Discontents*, edited by Carl Gillett and Barry Loewer (Cambridge University Press, Cambridge, U.K., 2001). Pp. 74-98.

proper function of the wing is to support flight depends on the assumptions that having wings is part of the causal explanation for flight, and that flight is part of the causal explanation for the successful survival and reproduction of birds, bats, insects, and so on.

However, as Peter van Inwagen and Trenton Merricks have argued,¹⁶ a consistent physicalist should reject the existence of macroscopic objects like wings. All the causal work supposedly to be done by wings is actually done by a large number of fundamental particles arranged wing-wise. Analogously, the macroscopic *property* of being arranged flight-wise or being arranged wing-wise do no causal-explanatory work.¹⁷ For a physicalist, all of the real explanatory work is done by simply aggregating the properties of a large number of particle-trajectories. Macroscopic properties like being wing-shaped or flying do not cut the world at its causal joints.¹⁸ They are, for the physicalist, grue-like properties, massively disjunctive, gerrymandered properties. They seem natural to us only from an anthropomorphic perspective. When we describe a bird as flying, we are thinking of it from the perspective of reverse engineering: we are imposing upon the bird a hypothetical design plan. We are projecting upon the bird the intentions that we would have if we were trying to design such a creature for the tasks of survival and reproduction. A physicalist cannot imagine that describing natural things in this way reveals genuine, mind-independent causal connections.

¹⁶ Trenton Merricks, *Objects and Persons* (Clarendon Press, Oxford, 2001).

¹⁷ Paul Humphreys was the first to recognize this implication of physicalism: “How Properties Emerge,” *Philosophy of Science*.64(1997):1-17.

¹⁸ The difference between “microscopic” physical properties and “macroscopic” physical properties has nothing to do with the number of particles involved. Corresponding to each macroscopic property (like *being arranged wing-wise* or *being in flight*) is a genuine micro-physical property describing the states and trajectories of each of the particles composing the relevant macro-objects. Even if Paul Humphreys is right that the individual microproperties “fuse” into a single holistic property when whole are formed, the resulting property of the whole will still be microphysical in my sense: something like an n -particle quantum state, for a very large n . Considered as a type, each properly macrophysical property (properties corresponding to gross, measurable and observable features of ordinary objects) corresponds to a wildly disjunctive microphysical property, disqualifying it from playing a genuinely explanatory role.

Of course, there is nothing wrong, from a physicalist point of view, in offering an “explanation” of a macroscopic event in terms of other macroscopic events. There is nothing contrary to physicalism in supposing that thrown baseballs do “explain” (in some legitimate sense) broken windows. The physicalist can take such “explanations” as convenient shorthand for some genuine causal explanation (of a kind that could be offered only by a Laplacian intellect). Presumably, whenever a good macroscopic “explanation” can be given, this quasi-explanation bears some analogy to the genuine, microphysical explanation of the event. Some of the physical quantities involved in the actual explanation will have some vague or approximate counterpart in the macroscopic quasi-explanation.

However, although such quasi-explanations are good enough for everyday life, they are not good enough for use in a metaphysical *theory* about the nature of mental states. If we are to suppose that we can capture the essence of mental states by Ramseyfying some functional or teleofunctional theory Ψ , then we must suppose that Ψ specifies, in metaphysically serious terms, the actual causal profile of the various mental states. This means that Ψ must specify the environmental inputs and outputs (such as tissue damage or aversive behavior) in exclusively microphysical terms. Ψ must be the kind of theory a Laplacian intellect would construct, reflecting bona fide causal explanations. However, in such a theory, there will be nothing that corresponds either to the mental states of folk psychology or the teleofunctional states of biology (including neurology), since these are specified in terms of gross, macroscopic inputs and outputs.

In fact, the number of functional states in a properly micro-physicalistic theory will be many orders of magnitude greater than the number of functional states in a corresponding theory Φ of biology, psychology, or any of the other special sciences. The latter sort of theory specifies the inputs and outputs in gross, macroscopic terms. The multiple realization problem that Functionalism was supposed to solve emerges again with respect to the environmental inputs and behavioral outputs: each macroscopically specified input or output can be realized by an astronomical number of distinct microphysically specified states. The special-scientific theory Φ will contain predicates specifying the relevant

input and output states in gross, macrophysical terms. Say that Φ contains n such specifications: $\phi_1, \phi_2, \dots, \phi_n$. A corresponding theory Ψ that specifies these inputs and outputs in exclusively microphysical terms will contain a large number of microphysically specified predicates for each corresponding input/output predicate in Φ : $\Psi_{1,1}, \Psi_{1,2}, \dots, \Psi_{1,m}, \Psi_{2,1}, \dots, \Psi_{2,m}, \dots, \Psi_{n,m}$.¹⁹ Since there are many more input and output states in Ψ than in Φ , there will be, correspondingly, a much larger number of internal states – states that will be replaced by second-order variables when the two theories are Ramseyfied. Thus, the microphysical theory Ψ will contain a much larger number of functional states than will be contained by the corresponding special-scientific (e.g., biological or psychological) theory Φ .

This poses a serious problem for any kind of Functionalism, since there seems to be little reason to think that the properly microphysicalistic theory Ψ of functional states will bear any resemblance to the theory Φ of any special science that is humanly discoverable. In particular, there is no reason to suppose that there will be a homomorphism from the states of the proper microphysical theory Ψ into those of any such special-scientific theory Φ . If we suppose, initially, that in the Ramseyfied version of the special-scientific theory Φ , each state-variable X_k corresponds to a set of variables Y_{k1}, \dots, Y_{km} in the Ramseyfied version of the microphysical theory Ψ , then it is very unlikely that the state-transition function for Φ matches exactly the state-transition function for Ψ , in the sense that whenever theory Φ specifies a transition from X_i through X_j , the microphysical theory Ψ specifies a transition from a state Y_{i1} (belonging to the set corresponding to X_i) to a state Y_{jk} (belonging to the set corresponding to X_j). The microphysical differences among the states associated with a functional-state for theory will undoubtedly result in

¹⁹ To be as charitable as possible to the Functionalist, I am assuming here an epistemic account of the vagueness of our macroscopic predicates, that is, that each predicate corresponds to a disjunction of a *precise* set of microphysical predicates. If we were to adopt instead a semantic theory of such vagueness (as I and most other philosophers would think is needed), then any Functionalist theory of mental states is yet one more step removed from reality. Few, if any, of our attributions of mental states would be “super-true”, true in all possible precisifications of a Ramseyfied folk psychology.

significant, gross differences downstream, given the extreme sensitivity of systems like the brain to small initial differences.²⁰

Thus, at best, we can hope that any special-scientific theory Φ stands in a relation of imperfect but nonetheless useful approximation to the Laplacian theory Ψ . In other words, the functional theories of the special sciences are at best useful fictions. However, this conclusion is devastating to the kind of etiological account of teleology under examination. If in fact, there are no states (not even highly disjunctive, gerrymandered states) corresponding to the second-order variables of a special-scientific theory, then it cannot be the case that some such state is caused to exist (is selected for by nature) in part because it bears certain causal relations to other states. Useful fictions cannot stand in genuine causal relations to other useful fictions. Another useful principle: *fictional entities cannot enter into real causal relations*. At best, we can say that it is useful to *pretend* that such causal relations hold. This means, however, that we cannot take teleofunctional theories realistically, and so we cannot take the etiological account of proper functions as establishing the real existence of such things.

This problem for the etiological account becomes even worse when we consider higher-order functions: teleofunctions whose proper function is specified in terms of teleofunctional inputs and outputs. For example, the capacity for deductive inference, such as the capacity to perform modus ponens inferences, is such a higher-order

²⁰There is considerable evidence for the hypothesis that the brain is a chaotic system, subject to the amplification of small initial differences: see Robert C. Bishop's 1999 Ph.D. dissertation at the University of Texas at Austin: "Chaotic Dynamics, Indeterminacy and Free Will"; Larry Vandervert, "Understanding Tomorrow's Mind: Advances in Chaos Theory, Quantum Theory and Consciousness in Psychology," *Journal of Mind and Behavior* 30(1997):25-35; David V. Newman's 1995 Ph.D. dissertation at the University of Texas at Austin, "Chaos and Consciousness"; Erol Basar, "Chaotic Dynamics and Resonance Phenomena in Brain Function," in *Chaos in Brain Function*, edited by Erol Basar (Springer-Verlag, Berlin, 1990), pp. 1-30; Jeffrey Foss, "Introduction to the Epistemology of the Brain: Indeterminacy, Micro-Specificity, Chaos and Openness," *Topoi* 11(1992):45-57; Christine A. Skarda and Walter J. Freeman, "How Brains make Chaos in order to Make Sense of the World," *Brain and Behavioral Sciences* 10(1987):161-173; Christine A. Skarda and Walter J. Freeman, "Chaos and the New Science of the Brain," *Concepts in Neuroscience* 1(1990):275-285.

teleofunction. The capacity to apply modus ponens involves a state that produces, in response to the presence of beliefs whose logical forms are P and *if P , then Q* , a belief of the form Q . Beliefs are themselves teleofunctional states of a certain kind (states whose proper function includes carrying the information that a particular kind of state of affairs hold). For the etiological account of teleology to apply to such higher-order functions, two things must hold: (1) there must be causal connections between, on the one hand, a pair of beliefs and a particular mechanism, and, on the other hand, an output belief, and (2) this causal connection must have contributed to the origin or perpetuation of that very mechanism. However, I have established above that the first-order teleofunctional states (including beliefs) are merely useful fictions. Therefore, the causal connections between beliefs, and the causal relevance of this connection to the existence of a cognitive mechanism (which, in turn, is itself a mere fiction), must be yet another layer of fictional representation. How do we decide when it is useful to posit such fictional causal connections? The usefulness of such a fiction, removed from reality by so many layers of additional fictions, would seem to have nothing to do with the actual causal connections holding in the world. Hence, the so-called etiological account of teleology, when combined with physicalism, collapses into the kind of vulgar instrumentalism (adopting the “design stance” or the “intentional stance”) advocated by Daniel Dennett.

To summarize: strict adherence to physicalism poses insuperable difficulties for a Teleofunctional version of Functionalism. It is essential to the etiological account of teleofunctions (as developed by Taylor, Wright, Millikan, Dretske et al.) not only that there be a causal connection between some structure and some piece of behavior or attunement to the environment, but also that this very causal connection be part of the causal explanation of the perpetual existence of that structure. So, for example, for birds’ wings to have flight as their proper function, it is necessary both that wings contribute causally to flight, and that this causal contribution itself be part of the causal explanation of the continued existence of winged birds. However, as we have seen, *flight* is not the sort of thing that can genuinely be the effect of any microphysically specified cause. At best, the notion that wings contribute to flying is a quasi-explanation, a rough-and-ready shorthand that corresponds, in each case, with some detailed causal explanation at the

microphysical level. In other words, the idea that wings contribute causally to something called “flight” is at best a useful fiction. But, it must be much more than a fiction if the etiological account of proper functions is to get off the ground (so to speak). Without a genuine causal explanation that links *having wings* with *flying*, it is impossible for wings to have flight as their proper function, since we the causal explanation of the continued existence of winged birds cannot depend on a causal connection that simply is not there to be depended on.²¹

Thus, except for microscopic functions, like hemoglobin's function of binding and releasing oxygen molecules, the teleofunctional account cannot account for biological proper functions, if physicalism is assumed. A fortiori, it cannot account for the mental functions of brain states.

Similarly, traditional Functionalism in general assumes that macroscopic inputs (sensible objects) and macroscopic outputs (gross behavioral motions) can be causally explanatory, assumptions that are incompatible with physicalism. As Merricks and Paul Humphreys have pointed out, all the considerations that motivate physicalism also motivate microphysicalism, the view that the microphysical world is causally closed. Macroscopic physical properties have no more causal efficacy than explicitly non-physical ones. Thus, to avoid the causal exclusion problem, a Functionalist theory would have to describe all the relevant outputs and inputs in purely microphysical terms. It is clearly impossible for such a theory to capture things even remotely analogous to mental states.

8. Too High a Price?

If mental causation really is incompatible with physicalism, then why not simply give up on mental causation? Isn't the price for commitment to a non-physicalist metaphysics simply too high?

²¹ For further objections to the etiological account of functions see Michael C. Rea, *World without Design: The Ontological Consequences of Naturalism* (Clarendon Press, Oxford, 2002), Chapter 5.

But what are the alternatives to mental causation? There are of course two: epiphenomenalism or eliminativism. Epiphenomenalism, the thesis that mental states are real but causally inert, poses insoluble problems for semantics and epistemology. How is it that we are able to acquire knowledge about mental states, or even so much as represent or refer to them, if there are no causal connections from mental states to our words or behavior? Mind-free zombies could, by hypothesis, behave exactly as we do, making the same assertions about their non-existent mental life that we do about our own.

Of course, both epiphenomenalism and eliminativism represent very radical departures from the conception of reality shared by all human beings from at least the dawn of history. Such a radical mutilation of our inherited web of belief is a course to be avoided, if at all possible. In addition, there is something self-defeating about a skepticism about the mind that is based on the results of modern science. The practice of science, no less than other human enterprises, is shot through with assumptions about the existence and efficacy of human minds. Science has always been a profoundly social activity. Scientists depend heavily upon fairly naïve, commonsense assumptions about human psychology whenever they rely upon the veracity of statements made in reputable journals and scientific meetings. If commonsense is wrong about the reality of the human mind and its activity in the world, then we lack any grounds for trusting in the deliverances of modern science, including those that would be needed to overthrow the reality of mental causation. A scientific eliminativism would be a self-defeating defeater of folk psychology.

In addition, unlike the eliminativism of Steven Stich or Paul or Patricia Churchland, the kind of eliminativism that would result from the exclusion problem would deny, not only that there are mental states as folk psychology conceives them, but that there are any states even remotely like mental states. The Stich or Churchland position would be better labeled “replacementism”, since they argue, not that there is nothing in nature at all like beliefs and other mental states, but only that the mental properties posited by folk

psychology should be replaced by syntactic or neurological properties playing analogous roles.

The argument I've sketched above would apply equally to the syntactic and neurological states favored by Stich and the Churchlands: only strictly microphysical states would be left. This kind of radical eliminativism is self-defeating in a very strong sense, since it implies the non-existence, not only of belief and assertion, but of anything remotely like belief or assertion. Such a radical eliminativist cannot coherently take himself to be asserting or advocating belief in eliminativism, or to be criticizing acceptance of its rivals. To adopt such a position is to pay a very heavy price indeed.

There remains only the option of instrumentalism: that mental states are merely useful fictions. Besides being scarcely credible, such fictionalism about the mental is incoherent. A useful fiction requires someone to understand and apply it. If the very existence of subjects of understanding and intentional use is part of the fiction, there is no one there to whom the fiction could be intelligible or useful.