

FUNCTIONALISM WITHOUT
PHYSICALISM: OUTLINE OF AN
EMERGENTIST PROGRAM

Robert C. Koons

November 22, 2002

Functionalism without Physicalism

Abstract

The historical association between functionalism and physicalism is not an unbreakable one. There are reasons for finding some version of a functional account of the mental attractive that are independent of the plausibility of physicalism. I develop a non-physicalist version of functionalism and explain how this model is able to secure genuine emergence of the mental, despite Kim's arguments that such emergence theories are incoherent. The kind of teleological emergence of the mental required by this model is in fact fully compatible with the best available interpretations of physics and does not simply repeat the mistakes of *vitalism*. In addition, this model of teleological, emergent causation provides an attractive account of free/libertarian agency.

1 Physicalism versus Emergence

Functionalism, as originally conceived, was a way of locating mental properties within a causally closed physical world. More recently, the Functionalist program has taken a turn toward teleology, incorporating a causal account of proper functions within a causal-role theory of mental states. However, doubts about the consistency of physicalism with genuine mental causation have persisted. The attempt to secure the causal efficacy of the mental has revived interest in alternatives to physicalism, including theories of the *emergence* of the mental.

The historical association of functionalism with physicalism has prevented the merger of the programs of functionalism and emergence. However, there are advantages to a Functionalist account, independent of its association with physicalism, that make a teleofunctionalist version of emergence quite attractive. In this paper, I will develop such a non-physicalist version of Functionalism ("Neo-Functionalism") and argue that this has a number of advantages as a model for emergence, meeting Kim's challenge to the coherency of emergent causation, and fitting emergence smoothly within the picture of the world sketched by the best interpretation of contemporary physics.

2 What's Wrong with Functionalism

In David K. Lewis's version of Functionalism,¹ 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 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

¹"Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50 (1972):249-258.

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, Millikan² 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.

Lewis/Ramsey Functionalism depends on the kind of view of mental causation expounded by Donald Davidson in “Mental Events”.³ 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 (i.e., instantiations of mental properties) themselves do not enter into causal explanations.

As Jaegwon Kim has pressed 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.⁴ This tension is especially acute when we consider the categories of responsibility and agency (as Hawthorne and Cover, O’Connor and William Hasker have recently argued).⁵

²D. 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).

³“Mental Events”, in *Essays on Actions and Events* (Clarendon Press, Oxford, 1980), pp. 207-27.

⁴*Supervenience and Mind* (Cambridge University Press, Cambridge, U. K., 1993); *Mind in a Physical World* (MIT Press, Cambridge, Mass., 1998).

⁵John Hawthorne and J. A. Cover, “Free Agency and Materialism”, in *Faith, Freedom and Rationality*, Jeff Jordan and Daniel Howard-Snyder, eds. (Rowman & Littlefield, Lanham,

However, Barry Loewer has recently pointed out,⁶ that 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 reducible to 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.

Such Humean and semi-Humean accounts of causation are unacceptable from a metaphysical point of view, as I have argued elsewhere.⁷ 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 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.

However, Humean and semi-Humean accounts do not respect this principle of intrinsicity, since whether one property-instantiation causally explains another depends, on these accounts, on remote facts about the correlations of these property, either throughout the actual world alone, or throughout the actual world and nearby possible worlds.

3 What May Be Right about Functionalism

Physicalism (the thesis that the physical realm is causally complete) is incompatible with the reality of mental causation. This incompatibility applies to all versions of physicalism, whether reductive or non-reductive. In particular, it applies to physicalistic Functionalism.

The incompatibility of Functionalism with mental causation depends upon the wedding of Functionalism with physicalism. Before we reject Functionalism altogether, we need to consider whether a divorce of Functionalism from

Md., 1996), pp. 55-69; Timothy O'Connor, "Causality, Mind and Free Will", in *Philosophical Perspectives 14: Action and Freedom*, edited by James Tomberlin (Blackwell, Cambridge, 2000); William Hasker, *The Emergent Self* (Cornell University, Ithaca, 1999).

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

⁷In "The Metaphysics of Mental Causation: Explanatory Exclusion Redux", submitted to *Philosophy and Phenomenological Research*.

physicalism is possible, and, if so, whether there is anything that might recommend to us a non-physicalist version of Functionalism, a position that I will call “Neo-Functionalism”.

Neo-Functionalism agrees with Functionalism in identifying mental states with higher-order physical states, that is, with the state of being in a physical state with a certain characteristic causal role. The form of Neo-Functionalism that I favor starts with the true theory $\Psi(X_1 \dots X_n)$ of the teleological proper functions of the human being (including a specification of the proper end or telos of human life). Each variable X_I corresponds to one of a human being’s proper function, including mental functions like modes of perception, inference or action. If X_k corresponds to a mental function, then Neo-Functionalism, just like Functionalism, identifies the corresponding mental state with the higher order property $\lambda y \exists X_1 \dots X_n (\Psi(X_1 \dots X_n) \& X_k(y))$.

There are two critical differences between Functionalism and Neo-Functionalism. First, as I have said above, Neo-Functionalism rejects the causal closure of the domain of first-order physical states. Neo-Functionalism embraces genuine, emergent downward causation, causation that makes an ineliminable difference to the objective chance of physical events. Second, unlike Functionalism, Neo-Functionalism is not committed to denying that mental states lack any intrinsic character. For Neo-Functionalism, higher-order, mental and other teleological states are first-class citizens of the ontology, not an ontological free lunch. These higher-order properties have instances in their own right – they are not merely instantiated by virtue of the instantiation of other, first-order properties. I will call such properties *first-class properties*. In other words, instantiations of first-class properties are first-class citizens of our ontology: each first-class property is instantiated either by distinctive tropes or by distinctive facta. First-class properties, unlike second-class ones, are not instantiated *by virtue of* the instantiation of other properties.

I see no reason for a Neo-Functionalist to accept Sydney Shoemaker’s thesis of *causal structuralism*,⁸ the thesis that the essence of a property consists wholly in its nomological/causal connections with other properties. If we reject causal structuralism, we must suppose that each first-class property has its own *quiddity* or *suchness* (analogous to the *haecceities* or *thisnesses* of concrete particulars). Just as an haecceity (like the property of being Socrates) identifies its bearer as a unique individual, so a unique quiddity is associated with each metaphysically basic, ontologically first-class property. The quiddity is the intrinsic, occurrent or qualitative aspect of a property.

For old-fashioned Functionalists, only first-order physical properties have quiddities: higher-order properties have only causal roles. For Neo-Functionalists, in contrast, some higher-order properties, including all mental ones, do have their own quiddities. In a paper in progress, entitled “Qualia are Quiddities”, I argue that we can identify the so-called phenomenal qualia, the “raw feel” in Wilfred Sellars’s phrase, of mental states with the corresponding quiddities. The

⁸Sydney Shoemaker, “Causality and Properties,” in Peter van Inwagen (ed.), *Time and Cause* (Dordrecht, Netherlands, D. Reidel, 1980), pp. 109-135.

intrinsic reddishness of a red sensation is simply the quiddity of the property being appeared to redly (in Chisholm’s language).

This account takes qualia seriously as inhabitants of the worlds, and yet provides grounds for rejecting the possibility of inverted (or otherwise scrambled) color spectra. The state of sensing a green object cannot bear the qualia of reddishness, since that state cannot, being the state it is, bear the quiddity of a distinct state. Two properties can no more exchange their quiddities than two people can exchange their haecceities. Bush cannot become Cheyney, nor vice versa. They can switch positions and even switch many of their occurrent properties, but they cannot literally switch identities. Similarly, the quale of reddishness is correlated as a matter of metaphysical necessity with the property of being appeared to redly. “Another” property couldn’t have bear that very same quale without being the property of being appeared to redly.

In later sections, I will sketch a version of Neo-Functionalism, but before embarking on that task, I want to consider the other question: is there any reason, apart from a commitment to physicalism, for finding a Neo-Functionalist account of the mind attractive? There are in fact two such reasons. First, Neo-Functionalism provides us with a plausible account of the essences of mental states. When we identify mental state-kinds by reference to certain central causal roles that these states play, we are doing more than *fixing the reference* (in Kripke’s sense) of the corresponding mentalistic terms. For example, the connection (in normal subjects and normal circumstances) between *pain* and *aversive behavior* is not a merely accidental one. It is of the essence of pain that it motivates aversion. Similarly, the normal connection between a squarish sense-impression and the presence, in the appropriate environmental location, of a squarish object, is essential. I agree with Brentano that intentionality is a mark of the mental, and I believe that intentionality involves causality. Hence, the characteristic causal role of a mental state is essential to it.

Moreover, Neo-Functionalism offers us an alternative to mysterianism (à la Colin McGinn) about the nature of mental properties. If we can identify mental states with certain instances (tropes or facta) of higher-order physical states (the existence of a physical states satisfying a certain causal role), then our metaphysics will be simpler than it would be under a more radically anti-reductionistic dualism.

4 Building a Teleological Version of Neo-Functionalism

Suppose that there exist tropes of higher-order states, tropes that are no longer thought of as part of an ontological free lunch. Instead, these event-aspects are posited as first-class citizens of our ontology. How could such things fit into a coherent picture of the causal history of physical events?

Kim’s arguments against mental causation depend on taking the determination relation that holds between a physical state and a supervening mental state as a quasi-causal kind of dependence, with the mental state quasi-causally dependent on its physical realization. Given this assumption, the mental state

seems doomed to irrelevancy, even apart from a prior commitment to physicalism, since it is hard to see how a higher-order state (the instantiation of an existential generalization) could have any effect on the subsequent course of physical events over and above the effects of the first-order state that makes the existential generalization true (by constituting an instance of the generalization).

However, Kim's assumption about the causal dependency of a generalization on its instance is not obviously correct. The way in which an instance of an existential generalization can be said to determine the existentially generalized fact is not a causal determination of any kind. Since the dependency in that direction is non-causal, we would not be introducing anything like a causal loop in supposing that the existentially generalized trope causes a trope of property that instantiates the existentially generalized trope. In other words, we can entertain seriously the possibility of downward causation from higher-order to lower-order properties.

It is plausible to suppose that the first-order trope (the trope of the first-order property instance) *necessitates* the second-order trope (the trope of the existentially generalized property, like a Functional property). However, necessitation is not the same thing as causation. In fact, in *Realism Regained*, I gave several arguments for thinking that it is effect-tropes that necessitate their causes, and not vice versa. If that is right, then it would be very natural to think of the higher-order trope as causally prior to the first-order one.

5 Securing the Efficacy of Higher-order States

Kim's argument for the explanatory exclusion of the mental by the physical depends on supposing that the domain of first-order tropes is causally complete, rendering any higher-order tropes epiphenomenal. To escape this exclusion, we must try to imagine what it would be like for a lower-order trope to occur because it is an instance of a higher-order property. I am going to assume an indeterministic and probabilistic form of causality, of the kind developed in Chapter 6 of *Realism Regained*. The model of higher-order causation that I will develop can be called teleological propensity enhancement, or TPE.

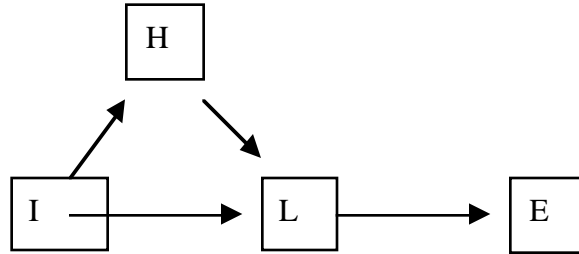
Let's consider a specific example in which there are four tropes (or facta). There is an initial trope *I*, an intermediate first-order trope *L*, a corresponding higher-order trope *H*, and a final trope *E*. Trope *E* constitutes an objectively desirable end-state, a telos for the system in question. Trope *L* constitutes the means, a physical mechanism, by which the end *E* is reached, and *H* corresponds to the Functional property of having a suitable first-order physical state capable of causing *E*. I am going to distinguish two probability functions, $Prob^0$ and $Prob$. $Prob^0$ represents the non-teleological or pre-teleological propensities for tropes to occur, while $Prob$ represents the all-things-considered propensity, one including any teleologically-grounded components. (Unlike standard conditional probability functions, I intend ' $Prob(A/B)$ ' and ' $Prob^0(A/B)$ ' to be undefined unless *A* is causally dependent on *B*.)

According to the TPE model, the posterior probability $Prob(E/I)$ will be higher than the prior probability $Prob^0(E/I)$, given three conditions: (i) $Prob^0(E/L)$ is relatively high (above some threshold δ), (ii) $Prob^0(L/I)$ is significant, non-negligible (above some threshold μ), and (iii) E represents an objectively good end-state (or one objectively good for the system or organism in question).⁹ In other words, state L is effective for producing the desirable state E , and the leap from I to L is not too great.

The higher-order trope H should then be thought of as an concrete instance of the following higher-order property:

$$\lambda y \exists X (Prob^0(E/X) > \delta \ \& \ Prob^0(X/I) > \mu \ \& \ X(y))$$

The TPE principle entails that the probability of the occurrence of H is higher, due to the objective goodness of E , then it would otherwise be. This entails that the probability of any first-order state verifying the second-order existential generalization included in H , including L , is also higher than it would otherwise be. In fact, the objective probability of L is higher *because* it is an instance of the existential generalization in H . We can capture this dependency by supposing that H is causally prior to L , which means that H causally explains the occurrence of its instance L . The direct causal connection from I to L does not render H redundant, since the ateleological propensity of L given I , $Prob^0(L/I)$, is lower than the all-things-consider propensity, $Prob(L/I)$, and the all-things-considered probability $Prob(L/I)$ is higher because H is instantiated.



H does not necessitate L , since there will ordinarily be other possible states L' , L'' , etc., similarly related to the existential generalization in of H , states with a non-zero probability given initial conditions I . In many cases, however, the conditional probability of L on H and I will be quite high, close to 1. L does necessitate H , since once the object in question realizes state L , given the probabilistic connections between I and L and between L and E , the object ipso facto realizes state H as well. However, necessitation is not causation: in fact, as I mentioned above, I hold that causes never necessitate their effects; so, H cannot be causally dependent on L .

⁹In *Realism Regained: An Exact Theory of Causation, Teleology and the Mind* (Oxford University Press, New York, 2000), I sketch an account (which I call “Aristotelian”) of objective goodness in terms of the existence of a system of harmoniously related proper functions. Something similar could be done here: condition (iii) could be cashed out in terms of the existence of a mutually supporting system of end-states meeting conditions (i) and (ii).

Let's consider a slightly more concrete application of this model. Let H be the teleofunctional property of having a desire for the formation of a friendship. Any physical realization of this higher-order property (in the form of a distribution of various connection-strengths throughout the brain's neural network) will increase the person's propensity to obtain an objectively good end E , that of friendship. If the brain's initial state I is one from which the physical propensity of reaching at least one neural realization of the desire for friendship is above the threshold μ , then the TPE model would predict that the probability of a transition from the initial state I to some neural realization of the desire is enhanced by the fact that that neural state is a realization of a desire for the good of friendship. When a neural state L results that realizes the desire for friendship, we should say that the neural state L was caused, in part, by the higher-order state H , that is, that the desire for friendship, which emerged from I , was in part the cause of its own neural realization L . We need to refer to the emergence of the desire for friendship in explaining why the probability of the occurrence of the neural realization of the desire was as high as it in fact was. Referring only to the physical propensities of I , the initial brain state, would, according to the model, not suffice. At the level of forces and energies, we should say that the occurrence of the desire for friendship and its effect upon the brain was associated with the activity of a non-physical force, and with the release of a corresponding form of potential energy.

However, the TPE model is still a form of Neo-Functionalism, since it predicts that the process by which the neural realization of the desire affects behavior and leads to an increased chance for obtaining the good of friendship is itself an entirely mechanical one, requiring no further invocation of mental forces. Thus, far from discouraging the search for neural mechanisms underlying intentional action, and far from being discouraged by the actual discovery of such mechanisms, the TPE model predicts that the characteristic effects of the realization of mental states will be fully explicable in physical terms.

6 Such Teleological Causality is Fully Compatible with Modern Physics

Many scientists and philosophers of science have assumed that the Galileo-Newton revolution in physics has consigned teleological explanation to the dustbin. However, this overlooks the continued vitality of teleological explanations in physics in the form of so-called 'variational principles', such as least action principles.¹⁰ Both classical and quantum mechanics can be formulated in terms of integral equations, which prescribe a path or trajectory that satisfies a holistic requirement, like the local minimization of action. In most cases, the same

¹⁰Wolfgang Yourgrau and Stanley Mandelstam, *Variational Principles in Dynamics and Quantum Theory* (Dover Publications, New York, 1979), pp. 19-23, 164-167; Cornelius Lanczos, *The Variational Principles of Mechanics* (4th edition, Dover Publications, New York, 1986), xxvii, 345-6; Robert Bruce Lindsay and Henry Morgenaw, *Foundations of Physics* (Dover Publications, New York, 1957), pp. 133-6.

physical theory can be case either in terms of differential equations (with associated notions like the composition of forces and the conservation of energy or momentum) or in terms of integral equations (corresponding to teleological explanation).¹¹ Many scientists assume that since the integral form of a theory can be transformed into a differential form, this means that the differential form represents the more fundamental mode of explanation. This is a non sequitur, however, since differential forms can similarly be transformed into integral forms. (In fact, there are pathological functions that can be integrated but not differentiated.)

Suppose we reverse the usual assumption and take the teleological explanation to be more fundamental. In one case after another, the teleological form of physical theory has proved to be both simpler and more fruitful than a teleological alternatives.¹² All of Newton's optics and mechanics can be derived from William Rowen Hamilton's formulation of least action. Both Einstein's equations of relativity and Schrödinger's equations for quantum mechanics can be derived from similar minimum action principles. If teleological explanation is the truly fundamental one, than the composition of forces and the conservation of energy and momentum would be, in reality, epiphenomenal in nature. Efficient explanation by means of differential equations would be a heuristically useful but ultimately fictional. Such a teleological recasting of modern physics (advocated most energetically by Max Planck¹³) would provide a physical model far friendlier to mental causation and agency than the usual metaphysical model, which treats efficient-causality at the level of forces as most fundamental. If teleology is primary in physics, then mental causation would be merely a large-scale version of a universal phenomenon.

On a teleological restructuring of physics, some higher-order states of particles would be causally efficacious, even apart from the cases of bio- or psycho-functional states. The state of moving along a trajectory that would, taken as whole, represent a local minimum of action, would be causally responsible for subsequent positions of the particle. The state of moving along such a trajectory is a higher-order, functional state: that state of moving to a position x such that the path from the particle's present position to x has the property of belonging to a trajectory that, taken as a whole, minimizes the quantity of action. It is this existentially-quantified, higher-order state that, from the teleological perspective, is primarily responsible for the particle's motions: the operations of the fundamental forces of physics simply supervenes on this still more fundamental level of causal explanation.

¹¹See Val Dusek, "Aristotle's Four Causes and Contemporary 'Newtonian' Dynamics", in *Aristotle and Contemporary Science*, vol. 2, D. Sfendoni-Mentzou, J. Harriangadi and D. M. Johnson, eds. (Peter Lang, New York, 2001), pp. 81-93.

¹²Jim Hall, "Least Action Hero", *Lingua Franca* 9 (October 1999): 68.

¹³Max Planck, "The Principle of Least Action", *A Survey of Physical Theory*, R. Jones and D. H. Williams, trans. (Dover Publications, New York, 1960), pp. 69-81; "Science and Faith", in *Scientific Autobiography and Other Papers*, W. H. Johnson, trans. (W. W. Norton & Co., New York, 1936), pp. 119-126.

7 TPE and Emergence

On the TPE model, there is a sense in which the mental *emerges from* the physical. The word *emergence* has a variety of meanings in philosophical contexts, but it is still a useful word, marking out those theories according to which mental properties are generated by configurations of physical particles in a law-like way. According to the TPE model, there are new fundamental forces at the level of the mental whose action is not supervenient on the actions of the familiar fundamental forces of physics (gravitation, electromagnetism and so on). This fits with John Stuart Mill's conception of emergent forces as responsible for *heteropathic* effects.

William Hasker has produced a useful taxonomy of notions of emergence, building on an earlier distinction of John Searle's: emergence 1a, emergence 1b, and emergence 2.¹⁴ The emergent-1a powers of a thing are fully explicable in microphysical terms. Emergent-a powers are emergent only in the sense that it is a conceptual truth that they cannot be manifested except by some relatively large collections of particles. So, for example, the power of being able to roll down a hill is a power that can be manifested only by a fairly large collection of particles arranged in a ball or a wheel. Nonetheless, the emergent-1a power of such a collection can be fully explained in terms of the operation of the normal fundamental forces of physics acting on each of the constituent particles.

An emergent-1b power is one whose action is not explicable in terms of microphysical forces. Novel fundamental forces and novel causal laws must be invoked to explain the operation of such forces. On the TPE model, mental powers are emergent-1b. Sydney Shoemaker¹⁵ makes a useful distinction between *micro-manifest* powers (such as those corresponding to the fundamental forces of physics) and emergent, *macro-manifest* powers. The mental powers of collections of particles are not explicable in terms of the micro-manifest powers of those particles.

An emergent-2 property is one that is caused by base properties but that causes effects that are not themselves caused by the base properties at all. Searle argued that the idea of an emergent-2 property is incoherent, since it would require a violation of the transitivity of causation. If base properties cause the emergent-2 properties, and causation is transitive, then the base properties must cause (indirectly) whatever the emergent-2 properties cause directly. I agree with Searle and would reject the thesis that mental properties are emergent-2. It is appropriate to say the the physical properties of the human behavior cause (indirectly) the human agent's intentional behavior, via causing the relevant mental properties. The potential for giving rise to mental activity is possessed by all physical particles, so no new, non-physical *entity* must be added to give rise to mental causation. However, genuine mental causation cannot be explained in terms of the operation of physical forces alone, in the sense of the fundamental forces that are manifest by the behavior of particles that are *not* part of the

¹⁴William Hasker, *The Emergent Self*, pp. 171-6; John Searle, *The Rediscovery of the Mind* (MIT Press, Cambridge, Mass., 1992), pp. 111-12.

¹⁵Sydney Shoemaker, "Kim on Emergence," *Philosophical Studies* 108(2002):53-63.

living body of a conscious agent (the “micro-manifest powers”).

However, it is possible that the emergence of novel powers does give rise to the existence of a new entities, namely, living organisms and human persons. A living organism is an entity that has both physical properties (like mass and location) and emergent biological powers (like digestion, metabolism and perception). Similarly, a human person is an entity with both physical and mental aspects. If we take a very skeptical conservative approach to the existence of *aggregates*, as Peter van Inwagen¹⁶ and Trenton Merricks¹⁷ have advocated, then we would deny that there is an entity corresponding to collections of particles that do not constitute a living organism or person. However, the teleological features of whole living bodies would justify the postulation of a new entity, the *organism*. The persistence of organisms and persons through time would coincide with the persistence-conditions of these emergent powers.

Do emergent properties supervene on their emergence-base? The answer to this question depends on whether we are thinking about synchronic or diachronic supervenience. If we are asking about synchronic supervenience, that is, whether a mental difference at some point in time would necessitate a physical difference at that same time, then, on the TPE account, the answer is clearly, Yes. Since mental properties are simply higher-order physical properties, a difference at the higher-order level necessitates a difference at the lower order. We cannot change the truth-value of an existentially quantified formula without changing the truth-value of at least one atomic formula.

What about diachronic supervenience? In other words, does fixing the base properties of the world at one point in time fix the propensities (the all-things-considered objective probabilities) for the occurrence of both base and emergent properties at all subsequent times? The answer to this question depends on what we include in the base. If we include all powers of particles, including the macro-manifest ones, then diachronic supervenience holds. However, if we include only the micro-manifest powers of particles, the operations of the four forces of physics, then diachronic supervenience fails. Were we to counterfactually subtract the power of particles to give rise to effects governed by principles of mental teleology, then the propensities for the occurrence of properties in the future, both base properties and emergent ones, would be different from what they actually are. What keeps the TPE model from being merely another form of physicalism is the irreducible mental nature of the teleological principles undergirding mental causation. Physical powers are distinguished from non-physical ones by virtue of the nature of the telos involved: physical forces correspond to least-action principles, mental forces to some kind of best-action or most-reasonable-action principles.

The TPE model has the resources for answering Jaegwon Kim’s overdetermination objection to the idea of emergence.¹⁸ Kim offers a two-pronged attack on downward causation by emergent properties. First, he argues that such down-

¹⁶Peter van Inwagen, “The Doctrine of Arbitrary Undetached Parts,” *Pacific Philosophical Quarterly* 62(1981):123-37.

¹⁷Trenton Merricks, *Objects and Persons* (Cambridge University Press, New York, 2001).

¹⁸Jaegwon Kim, “Making Sense of Emergence,” *Philosophical Studies* 105 (1999):1-34.

ward causation cannot be synchronic, because such synchronic dependence of some of the base properties on the emergent property would involve a vicious causal circularity, since the emergent property is synchronically dependent upon the base properties. An emergent property that had some simultaneous effect on the base properties would be to some extent self-caused. Second, Kim argues that downward causation by emergent properties cannot be diachronic either, since any diachronic effects of the emergent property would be over-determined, since they would also be caused by the coincident base properties.

I accept Kim's second argument. Since the causal powers of the emergent property are a proper subset of the causal powers of the more specific base properties, it would be redundant to attribute diachronic causal efficacy to the emergent property itself. Any causal efficacy by the emergent property is in effect preempted and superseded by the greater efficacy of its corresponding base properties. However, I reject Kim's first argument, since it depends on the faulty assumption that emergent properties are synchronically dependent on their base properties.

It is true, as I have admitted above, that emergent properties are synchronically supervenient on the base properties, but supervenience does not entail any kind of dependence, causal or otherwise. According to the TPE model, it is the higher-order, emergent property that causes its own lower-order instances, not the other way around. In fact, what happens is this: the realization of an existentially generalized property causes the realization of one of its instances. It is true that the effect in this case necessitates its cause (in fact, the necessity is a logical one), but that is no objection to the account. I have argued (in *Realism Regained*) that token-effects *always* necessitate their token-causes. The cause and effect are logically separable on the TPE account, since the higher-order property-instance does not necessitate any particular one of its possible instances. There will always be, in each situation, several alternative instances of the existential generalization that are possible effects of the higher-order cause.

If the lower-order state is causally dependent (synchronically) on the higher-order state, how is the TPE model still a model of *emergence*? In what sense does the higher-order state emerge from lower-order ones? The answer is simply, diachronically. The emergent state is nomologically dependent on temporally prior lower-order states. Thus, the TPE view of downward causation is opposite to the conventional picture: the higher-order state is diachronically dependent on prior lower-order states, and the lower-order states are synchronically dependent on the higher-order state. Such a reversal of the traditional picture is needed to meet Kim's challenge.

8 TPE as a Model of Agentive Causation

The TPE model provides an account of mental causation that fits closely many of the features of libertarian agency (as developed most recently by Robert Kane). It also represents a metaphysical model that shares many of the desirable

features of “agent causation” as proposed by Timothy O’Connor¹⁹ or William Hasker.²⁰ The TPE model depends upon a probabilistic account of causation, excluding the possibility of determinism. Hence, the TPE model is squarely within the incompatibilist/libertarian camp. Moreover, it is emergent, in the sense that the occurrence of teleological causation depends on the agent’s body being in an appropriate physical state (in the example above, a state I with the appropriate propensities). This means that TPE avoids the problem of anchoring the interaction between a particular soul and a particular body that bedevils many versions of substance dualism.

The TPE model would seem to occupy a position intermediate between the causal indeterminism of Kane and the agent-causationism of O’Connor. The emergence of new forces and energies at the psychological level would seem to be an “extra factor” of the sort that Kane has tried to avoid.²¹ That is, the TPE model involves a clean break from physicalism in a way that Kane’s model does not. However, Kane is increasingly sympathetic to a strongly emergentist approach, and also to allowing for a holistic, top-down kind of substance causation.²² And, like Kane’s model, the TPE model seeks to make human action scientifically explicable, as part of a unified model of causation. The TPE model does not posit a *sui generis* form of *agent causation* at the level of the human person radically different from the *event causation* that holds throughout the rest of the natural world. There is no incompatibility between recognizing the indispensable causal role of the whole human person as a continuing substance and analyzing agentive causation in terms of events and processes. As Kane puts it,

A continuing substance does not absent the ontological stage because we describe its continuing existence – its *life*, if it is a living thing – including its capacities and their exercise, in terms of states of affairs, events and processes involving it.

The TPE model involves a clearly holistic mode of causation. The end-state in question is typically a state of the entire organism. Like all teleological explanation, it is top-down rather than bottom-up. It avoids the threat to human agency that any atomistic, efficient-causal model poses. Such atomistic models make sub-personal factors bear all of the explanatory burden, making the person as such causally redundant. This is a problem with causal indeterministic models like that of Kane’s *The Significance of Free Will*, in which the “freedom” of the human person seems to merely supervene on independently occurring quantum events at the micro level. TPE dispels any such worries about the causal irrelevancy of the whole human person.

¹⁹Timothy O’Connor, *Persons and Causes: The Metaphysics of Free Will* (Oxford University Press, New York, 2000).

²⁰William Hasker, *The Emergent Self* (Cornell University, Ithaca, 1999).

²¹Robert Kane, “Free Will: New Directions for an Ancient Problem,” in Robert Kane (ed.), *Free Will* (Blackwell, Malden, Mass., 2002), p. 224.

²²*Ibid.*, pp. 228, 241.

Moreover, the TPE model is essentially ends-oriented. In the case of deliberate human action, this fact explains why the exercise of agency is inseparable from some kind of intention or purpose. One of the weaknesses of standard agent-causation models is that the connection between agency and intentions is posited as a kind of inexplicable brute fact.

Finally, the TPE model promises to cast light on the nature and significance of deliberation. Deliberation can be thought of as a process by which the human organism is moved into states in which new forms of teleological causation are triggered, as the objective probabilities of new and better ends are moved across the critical thresholds.

9 Dealing with the *Mind* Arguments

In a series of articles in the journal *Mind*,²³ several arguments were put forward (which Peter van Inwagen has collectively labeled ‘the *Mind* argument’) for the conclusion that it is indeterminism that is incompatible with responsibility, since it weakens the connection between an agent’s character and motivation on the one hand and his supposedly free actions on the other. The *Mind* argument can be seen as raising a family of related issues: the blind luck issue, the ownership issue, the character issue, and the deliberation issue. First, if my free actions are produced with a determinate probability, isn’t it a matter of dumb, brute luck (either good or bad) that I act as I do, as isn’t such blind luck incompatible with the action really being up to me? Second, if my actions are not causally explained by my character, beliefs, desires, pro-attitudes and so on, in what sense can my actions be thought of as truly mine? What links these undetermined actions to me as a responsible party? Third, evaluations of individual actions always take place within the context of an evaluation of the agent’s character. Responsibility is heightened, and not reduced, when we can see an action as an expression of a stable character, a set of virtues and vices. Fourth, indeterminism threatens to make the action of deliberation pointless, since indeterminism seems to introduce a causal gap between the deliverances of deliberation and the consequent free action.

The TPE model offers some hope for de-fanging these objections. The inherently teleological, purposive nature of TPE distinguishes free action from the case of blind, purposeless chance. The holistic and emergent character of teleological causation provides the basis for assigning ownership of the action to a particular human person. The ends invoked in such explanations may be biographical and characterological in nature, making the agent’s memory of his own past character directly relevant to determining the objective probabilities of various possible present actions. As I explained above, deliberation is fruitful because it has a tendency to raise the probability of higher and better ends.

Peter van Inwagen has developed a variation on the *Mind* argument known as

²³The first of these was R. E. Hobart, “Free Will as Involving Determination and Inconceivable without it,” *Mind* 169(1934):1-27.

the *rollback argument*.²⁴ We are to imagine some putatively free, undetermined human action, such as Susan's accepting a bribe. We then imagine the history of the world rolled back to a point in time just prior to Susan's action, and allow history to be replayed. This is repeated over and over again, resulting in a very long series of causally unrelated re-enactments. Since Susan's action is undetermined, it would seem that sometimes she takes the bribe and other times she refuses it. Van Inwagen claims that probability that the frequency of Susan's bribe-taking choices approaches a limit itself approaches *one* as the series becomes longer and longer. In reaching this conclusion, van Inwagen would seem to be relying on de Finetti's theorem. However, the applicability of de Finetti's theorem to this case depends on the assumption that any two series of Susan's choices that are permutations of each other (sharing the same frequency of bribe and non-bribe choices, albeit in different orders) have the same probability. One who wanted to deny that there is a well-defined probability of Susan's taking a bribe in a single case should also deny that there are well-defined probabilities for rollback series of Susan's choices, making de Finetti's theorem irrelevant.

However, there is some reason for agreeing with van Inwagen that there is an objective probability, in a single case, of Susan's taking a bribe. If robust mental causation is to be integrated into a unified picture of the natural world, then we will have to posit the existence of fundamental mental forces and potential energies, enabling us to preserve the energy conservation law. However, if there is no well-defined probability of the action of the mental, agency-related fundamental force, then it may be problematic to posit a corresponding form of potential energy, as David Papineau has recently argued.²⁵ I don't find Papineau's argument entirely compelling: it would seem that there could be laws specifying the maximum mental force that could be exerted by a particular arrangement of particles, without specifying a precise probability that this force would be in fact be exerted in various circumstances. There could be a definite quantity of potential energy available to mental agency in each situation, but perhaps only an interval-valued probability function detailing how that energy would be transformed into kinetic or chemical energy in each possible situation.

In any case, there would seem to be two forms of libertarian theory: chance and no-chance versions, a distinction that may be more significant than the usual contrast between agent-causation and other theories. A no-chance version of libertarianism, that seeks to resist van Inwagen's rollback argument by denying the existence of precise objective probabilities, would represent a much more significant separation of human agency from the rest of nature, as currently understood.

The TPE model as I have sketched it is clearly on the chance side of this contrast. I would argue that we find van Inwagen's rollback picture disturbing only because we assume that all chance can be modelled by the action of blind

²⁴Peter van Inwagen, "The Mystery of Metaphysical Freedom," in Robert Kane (ed.), *Free Will* (Blackwell, Malden, Mass., 2002), pp. 189-195.

²⁵David Papineau, "The Road to Physicalism," in Carl Gillett and Barry Loewer (eds.), *Physicalism and its Discontents* (Cambridge University Press, Cambridge, U. K., 2001), pp. 25-26.

and impersonal physical mechanisms, like tossing a coin or a die. However, the TPE model posits a sui generis form of chance at the level of mental causation, an essentially thoughtful and purposive form of chance. Once this is taken into account, I feel little inclination to deny that Susan would be fully responsible for her action on each and every occasion, even if the frequency would certainly approach a definite limit in the hypothetical long run. This limit reflects Susan's propensity to exercise her freedom in one way rather than another: I can't see how the existence of such a propensity deprives her of genuine freedom. It is still Susan, as an intelligent and purposive human agent, who makes the difference in each case: her actions are not finally explicable in terms of the propensity of her sub-personal parts and their arrangement.

10 The Objection to “Vitalism”

Much of the motivation for an allegiance to physicalism in philosophy derives from a sense that “vitalism” has been tried and found wanting. Certainly, some of the historical arguments for vitalism, such as those of Hans Driesch, were misplaced.²⁶ Driesch depended on arguments from mechanical impossibility. He claimed that certain operations of living things, such as reproduction or memory, are impossible to achieve by purely mechanical means. We now, with our knowledge of the DNA double helix and of computer architecture, have good reason to think that Driesch was just wrong. However, it would be a mistake to think that the failure of the Drieschian program had settled once and for all the question of whether a scientifically viable case for vitalism might be found.

We haven't found evidence for the existence of vital forces in places where Driesch would have expected we would, but we may simply have been looking in the wrong places. The TPE model of teleology, in sharp contrast to Driesch, predicts that we will find mechanical explanations for all biological and mental functions. It is the very possibility (with sufficient probability) of such mechanisms that would trigger teleological causality at the biological or psychological levels. What TPE would predict is that the probability of the initial formation (and perhaps the subsequent stability) of such mechanisms would be found to be higher than could be explained in purely mechanical terms. Teleological causality at the biological or psychological levels will become evident only when we are able to analyze with a high degree of precision the behavior of highly complex systems. Modern computers, with the capacity of modeling the behavior of systems involving millions or even billions of atoms, might provide the means for rigorous testing of various TPE hypotheses.

If I am right in thinking that the teleological mode of explanation is primary, and that explanations in terms of forces and potential energy are really epiphenomenal, then it may be premature to look for evidences of vital forces or vital potential energies. We need first a better idea what exactly are the biological

²⁶Hans Driesch, *The History and Theory of Vitalism*, C. K. Odgen, trans. (Macmillan & co., London, 1914).

and psychological analogues to the minimization of action at the level of mechanics. Once we have a precise version of the relevant teleological principles, we should be able to work out a corresponding theory of vital and mental forces and potential energies, much as David Bohm's mechanics converted quantum theory into the form of laws of efficient causality, with a corresponding form of quantum potential.

University of Texas at Austin