© 1998 Thomas W. Polger

Escaping The Epiphenomenal Trap

Thomas W. Polger
Department of Philosophy, Duke University
Box 90743, Durham, North Carolina 27708, USA

twp2@duke.edu
voice: 919.660.3065
fax: 919.660.3060

Escaping The Epiphenomenal Trap

I describe a feature of the debate between Functionalists and Anti-Functionalists in philosophy of mind that I call *The Epiphenomenal Trap*. I argue that the dialectic is a trap because neither side can resolve the central metaphysical issue as it has been put. That is because the debate typically trades in *possible explanations*. So long as Functionalists and Anti-Functionalists continue to debate whether functionalist explanations are *possible*, the central metaphysical issue cannot be resolved.

I examine what it is about the structure of functionalism that has persuaded many philosophers on both sides to take seriously possible functionalist explanations. I argue that possible explanations enter the debate due to a confusion over the metaphysical commitments of versions of functionalism.

It is in the interest of both sides to recognize and avoid The Epiphenomenal Trap, in order to address the metaphysical question. Thus, I urge that the debate between Functionalists and Anti-Functionalists be recast.

Escaping The Epiphenomenal Trap

I place my hand in a fire. The fire is hot. My hand hurts. The pain causes me to remove my hand from the fire. Pain has the function of causing persons to remove their hands from fire and other sources of injury. Functionalists and Anti-Functionalists alike could assent to this story. But the story is silent on the question of whether pain itself is to be given a purely functional characterization. Are the qualitative properties of conscious states really just properties of their functional organization, such that any way of having that organization is a way of *being a conscious state*?. Functionalists and Anti-Functionalists disagree about whether being a pain state is itself to be understood as being a functional state.

Functionalism

Functionalism is a theory about the metaphysics of mental states, "The diverse views gathered under the functionalist umbrella all agree that psychological and mental kinds are functional kinds." (Van Gulick 1988: 152) Functionalist theories of mind share a general form:

F. To be a mental state *M* of system *S* is to instantiate the functional role *F(m)* in the overall functional structure of *S*.

Being a mental state is a matter of *instantiating* a functional role. No other factors are relevant to being a mental state.

This is what is at issue between the Functionalist and the Anti-Functionalist. Are conscious state kinds functional kinds? The Functionalist says, yes; the Anti-Functionalist says, no.

Computer programs seem like good candidates for functional kinds. All there is to realizing a program is being a thing that has a certain functional organization. Functionalism in the philosophy of mind was inspired by computational theory; it holds that minds, like computers, are functional kinds.

Carburetors are candidates for functional kinds, metaphysical kinds whose essence is their function. One might think that anything that does what a carburetor does, anything that performs the function of mixing air and fuel, is a carburetor. After all, there are many ways of being a carburetor; that is, carburetors are *multiply instantiable*. It doesn't matter what company makes a carburetor, they all do the same thing—they are all carburetors. Automotive parts are classic examples of things that are sometimes thought to be functional kinds. But carburetors are not functional kinds. Consider: Carburetors control the mixture of air and fuel in the combustion chamber of internal combustion engines. The thing that performs the air-fuel-mixing function in a

1968 Mustang is a carburetor. But the thing that performs the air-fuel-mixing function in a 1996 Mustang is not a carburetor, it is a fuel-injector. Carburetors are not a functional kind because the carburetion function can be done by things that are not carburetors; it can be done, for example, by fuel injectors.[1]

The Epiphenomenal Trap

There is a line of discourse concerning the metaphysics of mind that is important enough to deserve a name: *The Epiphenomenal Trap*. My aim is to sketch the general shape of the Trap and suggest how it should be avoided. The Trap is subtle and I have no delusions that in the end I will have characterized it precisely, but I hope to direct those who are struggling with the Trap towards the escape routes, even if I have not fully opened them.

The Epiphenomenal Trap is not an argument. It is a feature of the philosophical discourse as it has been played-out for the past thirty years or so. The dialectic is between those philosophers who offer functionalist theories of consciousness and those who deny that the qualitative aspects of consciousness can be captured functionally.

Anti-Functionalist:	How could functionalism possibly account for consciousness?
Functionalist:	It is possible to give a functional theory of mind using my theory, *T* according to which consciousness is [*reduced-to* or *eliminated-in-favor-of* or i*dentified-with*] a non-conscious states *S* that plays functional role *F*. And *S*'s capacity to *F* is understandable entirely in terms of other functions or of entities and properties that are recognized by the physical sciences.
Anti-Functionalist:	No, it is not possible to give such an explanation. *T*, although it would account for many capacities (were it correct), cannot in fact account for capacity *A*. Conscious states are kinds that cannot be functionally analyzed, and some conscious state is necessary for *A*.
Functionalist 1:	*A* is important. Here's how *T* could account for *A*.
	Or,

[1] This example demonstrates the important and oft overlooked truth that things can be multiply instantiable without being functional kinds.

Functionalist 2: *A* is important, and *T* could not account for *A* as it stands. But with one small change we can change *T* into T_2; and T_2 could account for *A*. Still there is no irreducible phenomenal consciousness.

Or,

Functionalist 3: *A* is not important to understanding the mind. *A* does not exist.

Anti-Functionalist 1: *A* exists and it *is* important. Furthermore, neither *T* nor T_2 accounts for *A* because theories of that type treat the phenomenal *properties* necessary for *A*-ing as reducible to the *capacities* of other states or entities.

Or,

Anti-Functionalist 2: OK. Maybe you could account for *A*. But what about *B*....

This pattern has been continuing for years, each side raising the bar on the other, and it shows no sign of resolution. The Functionalists claim that their theories are gaining ground, and their opponents' conceding ground. The Anti-Functionalists claim that the Functionalists are simply missing the point.

What is an Anti-Functionalist to do? The options are:

a. Try again. Find a capacity *Q* for which consciousness is necessary.

b. Admit that consciousness, its phenomenal aspect at least, is eliminated.

c. Identify consciousness with a functional component of the physical stuff; consciousness would then have causal powers.

d. Insist that consciousness, from the phenomenal point of view, is just what we've always thought, but concede that it is an epiphenomena.

To the Anti-Functionalist, these options offer little solace.

Option (b) is eliminativism: there is no such phenomena as consciousness. I am not going to discuss reasons to reject eliminativism in this essay because I am primarily concerned with the dialectic between Anti-Functionalists and those Functionalists who are interested in explaining the phenomena of conscious experience, those for whom eliminativism is not acceptable.

Option (a) is the most frequently taken by Anti-Functionalists. Since the Functionalist holds that all mental kinds are functional kinds, the Anti-Functionalist need only be right about one capacity *Q* to show that functionalism is false. But it has proven extremely difficult to find some

capacity for which a kind of non-functional phenomenal consciousness is necessary. (This line of response is complicated by the fact that the Functionalist typically conceives of qualitative experience as a capacity, whereas the Anti-Functionalist conceives of qualitative experience as a property.)

Option (c) is functionalism. It identifies being conscious with instantiating a functional role. From the Anti-Functionalist point of view, functionalism has been thought to entail epiphenomenalism. If being a conscious state is a matter of instantiating a functional role, and instantiating a functional is a matter of standing in certain relations, then instantiating a functional role doesn't change any of the non-functional properties of a state. Being in a functional role doesn't seem to add any new properties to a state. In particular, it doesn't change the *causal* properties of a state. No state can have causal properties in virtue of being a functional state. If consciousness is functional, then consciousness cannot have causal efficacy *qua* consciousness. So the Anti-Functionalist is apt to see Functionalism as rendering consciousness epiphenomenal.[2]

Option (d) is explicit endorsement of epiphenomenalism. It purchases the phenomenology of consciousness by renouncing its efficacy and embracing its epiphenomenal status. Epiphenomenalism is undesirable because the Functionalists and Anti-Functionalists who are interested in explaining consciousness share the intuition that conscious states can have causal efficacy. As in the story with which I began, conscious states such as pain cause us to do things.

The Anti-Functionalist is faced with three choices: save the phenomenal aspect of consciousness at the cost of epiphenomenalism, save the efficacy of consciousness at the cost of the phenomenal aspect (eliminativism), or keep searching. This is one half of the Epiphenomenal Trap. If we assume that both sides are genuinely interested in explaining the phenomena of consciousness (i.e., if we discard eliminativism as an acceptable solution) then, from the Anti-Functionalist point of view, all available resolutions of the discourse end in epiphenomenalism. Hence, the *Epiphenomenal* Trap. The only non-epiphenomenal option is non-resolution—continue the discourse. Neither side is happy with that "solution", for there seems to have been no substantial progress for the past twenty years. But given the specter of epiphenomenalism, option (a) is clearly preferable to the Anti-Functionalist. Once more unto the breach!

The Functionalist Model of Explanation

Let's pause for a moment. "This is no trap," a Functionalist might reply, "it is simply the possibility space. The fact that it doesn't contain options that are to the Anti-Functionalists' liking

[2] Although to Anti-Functionalists the distinction breaks down, they must admit at least that the Functionalist follows a different strategy than the overt eliminativist would take.

doesn't make it a trap." And what's wrong with continuing to look for that capacity for which consciousness is necessary?

To be sure, The Epiphenomenal Trap is not a trap in the sense that one side is misleading the other, or that someone has an ace up their philosophical sleeve. The reason the Epiphenomenal Trap is a trap is that the Functionalists' own theory does not entitle them to argue for their metaphysical claim from the position that it is possible to give a functional account of consciousness.

The notion of a *possible explanation* is an odd one, to be sure. Philosophers of biology find it useful to talk about *how-possibly explanations* (Brandon 1990: § 5.3) to discuss conjectures about evolutionary histories, and I take it that those are a species of *possible explanation*. Possible explanations tell how something could *possibly* have happened, or how it could *possibly* be explained. Possibly explanations don't usually carry much weight. Generally one would like to be more confident in an explanation of a phenomena. One would like to think that the account was an explanation of how things are *actually*; at least an explanation ought to be thought to be likely, or in some way warranted. What it takes to turn a possible explanation into an explanation is evidence. Possible explanations without evidence are stories.[3] The availability of possible functional explanations is compatible with the falsity of functionalism.

It's not entirely the Functionalists' fault that they argue from this stance. Anti-Functionalists have encouraged it by demanding to know "how could functionalism possibly...?" But if the Functionalist is allowed to argue from that position, the debate is all but lost. And that's the trouble. It's all *but* lost, yet not lost; if we agree to argue about possible explanations the debate will forever be at a stalemate.

The problem is with the way in which the entire exchange is framed. When the Anti-Functionalist demands to know how functionalism could *possibly* account for human consciousness, the Functionalist is happy to oblige. For one thing, since no *complete* functional explanation of human consciousness has ever been formulated, the Functionalist has no *actual* explanation to offer. Thus it is fortunate that only possible explanations are demanded.

Second, possible explanations may include assumptions that actual explanations cannot. The Functionalist is entirely justified in beginning the possible explanation by *positing that conscious state kinds are functional kinds*. It is not an objection to Functionalists' possible explanations that consciousness is not in fact or may not be a functional category. After all, what was requested was a *possible* explanation; and it is certainly true that *if* conscious kinds are functional kinds, *then* functionalism is true.

[3] Gould and Lewontin (1978) call such stories about evolution "just so stories."

But whether conscious kinds are functional kinds is just what is in question. The Anti-Functionalist is right to be suspicious of possible explanations that are conditioned on mental kinds being functional kinds. The Anti-Functionalists should not keep searching for that capacity for which consciousness is necessary because they cannot win if the Functionalist is entitled to explanations that presume the functionalist metaphysic.

No merely possible functionalist explanation of overall human capacities poses any threat to the Anti-Functionalists' metaphysic, nor establishes the Functionalists'. And that is why The Epiphenomenal Trap is not only a trap for the Anti-Functionalist; it's a trap for the Functionalist, as well.

Functionalists have been satisfied to be caught-up in this way because they think they are at an advantage, whereas the Anti-Functionalists have felt the bite of the Trap because they have felt themselves to be on the defensive. But the Epiphenomenal Trap is a trap because *neither* side can settle the matter from this juncture; they are stuck together. As long as the debate centers around whether there is a possible functionalist explanation of consciousness, no progress can be made. Possible functional explanations go exactly no distance toward resolving the metaphysical question about the nature of conscious states. None. That is what makes the Epiphenomenal Trap a trap.[4]

If I gave you a possible explanation of how your car worked—one that, say, posited a super-strong super-fast platypus turning the drive shaft—you would not think that I had threatened your previously held beliefs about automobile engines. Yet the Anti-Functionalists have received the Functionalists' possible explanations with a great deal of consternation. Indeed, as we have seen, Anti-Functionalists have desperately tried to show that no functionalist explanation of consciousness is possible. And that opens the line of discourse that is The Epiphenomenal Trap. There must be something special about the structure of functionalism that persuades some philosophers to think that functionalist possible explanations have more power than other sorts of possible explanation.

[4] It might be objected I am misunderstanding what the Functionalist is providing. Functionalists are not just giving possible functional accounts of overall capacities of the system, they are giving (as the Anti-Functionalist demands) possible explanations for how consciousness itself—the qualitative aspects—could be understood in functional terms. That is so. But to do so the Functionalist conceives of the qualitative aspects of mental states as capacities. That is to say, functionalists give possible explanations of how consciousness could be explained if it were a capacity. And as we have seen, no merely possible explanation will budge the dispute.

The Structure of Functionalism

Functionalism is based on a distinctive model of explanation. First, figure out how a cognitive capacity works. When you have a functional analysis of the capacity, then anything that instantiates that functional structure *is*, functionally speaking, that kind of thing. Ideally, functionalism moves from *explanation* to *essence*; from what a kind of thing *does* to what kind of thing it *is*. Functionalism makes it a logical truth that the occupant of a functional role is a thing of a functional kind specified by reference to the role. Being a conscious state is identical to instantiating a functional role.

The relationship of *instantiation*, or *realization*, is fundamental to functionalism. (In the literature regarding functionalism as a philosophy of mind, 'instantiate' and 'realize' appear to be used interchangeably.) Functionalists distinguish between a functional role and the occupant of the role.[5] Instantiation is the relationship between role and occupant; occupants instantiate roles.

What relation is instantiation? Not identity; and not supervenience. What is it to instantiate a functional role? According to functionalists, instantiating a functional role is a matter of having a function. Functionalism makes claims about occupant states in virtue of the functional roles that they instantiate, in virtue of the functions that they have.

But what is it to have a function? There's the rub. There are various notions of function, and thus various notions of what it is to have a function. Most functionalist philosophers of mind seem to have accepted that various notions of function simply fill-in the details of a general "amorphous" (Van Gulick 1988) notion, the basic structure of which each particular version shares. This has been an error. Available notions of function are quite disparate; what it is to instantiate a function depends greatly on what kind of function is to be instantiated. If we look with care at how an early version of functionalism makes sense of the instantiation relation, we will see why one might think that possible explanations should carry weight—and why, in fact, they do not.

The original version of functionalism is computational functionalism. This is the sort of functionalism championed for some time by Putnam (e.g., Putnam 1960), inspired by insights by the likes of Turing. William Lycan characterizes the view thus:

> F_C: To be a mental state M is to realize or instantiate a machine program P and be in functional state S relative to P. (Lycan, 1987: 8)

Although this is a view that has been abandoned by many contemporary functionalists, it is important to us because functionalists have tended to assume that other versions of functionalism are variations on this basic structure.

[5] See Lycan (1987: 37) for a moving statement of the functionalist dedication to this distinction.

According to computational functionalism, instantiation is to be understood abstractly. Being a machine is having states in relations that correspond to the abstract input-output relations described by a machine table. Realizing a functional state is having certain formal relations to other states, and to the inputs and outputs of a machine. Being a computing machine, instantiating a computational function, is a matter of having these relations. Since all there is to instantiating a function is having certain abstract relations, anything that can stand in those relations can instantiate the function (multiple realizability.)

To show that a system instantiates a machine one need only show that it could be described by a machine table. Anything that can be described by an abstract function is a computing machine, on this view. Only one such functional description is needed to show that something can be described as instantiating a machine—to show that it *is* a machine.

This is how possible explanations enter the debate. Ideally the functionalist model of explanation reasons from explanation to essence. But to show that the mind is a computing machine it is only necessary that it have relations describable by *some* abstract function. That is, to show that the mind is a computing machine it is only necessary that it be *possible* to describe it in terms of an abstract input-output relation—it is only necessary to give a *possible* explanation. Understanding functions as abstract relations, if it is possible to describe the mind as a computing machine then *it is a computing machine.*

On this view of computational functionalism, it need only be possible to describe the cognitive system in a certain abstract way, for all there is to instantiating an abstract function is *being so describable.* The abstract functional description need not be the only or even the most useful description, just one possible description of the system. Thus, the functionalist model of explanation, when understood as dealing with abstract functional relations, seems to license reasoning from *possible* explanation to essence. That is why the Functionalists thought it important to show that there is a possible functionalist explanation of consciousness; and why the Anti-Functionalists thought it important to show that no such explanation was possible.

This picture of computational functionalism is the not very pretty one that Anti-Functionalists have faced. The only thing to say about it is that it is the wrong picture: computational functionalism does not ascribe abstract functions. The relations that machine tables describe are not formal or mathematical relations. Realizing a machine is not a matter of having an abstract relational structure; it is a matter of having a specific kind of relational structure—one involving causal relations.

Computational functions, it turns out, are a kind of *causal role function*.[6] Causal role functions are the effects of a thing that play a causal role in an explanation of an overall capacity of a

[6] Robert Cummins has given a formalization of this notion of function.:

containing system. Capacities of a containing system are explained in terms of capacities of components of the system in Cummins' "functional analysis" strategy of explanation. Functions are those capacities that are appealed to in such an explanation. (Cummins 1975)

Escaping The Epiphenomenal Trap

When it comes to giving causal role functional explanations, possible explanations won't do the trick. Causal role functions are *actual* effects that play explanatory roles. It's not enough that it be *possible* that a car be a system describable in terms of the causal role function in which a platypus is the source of locomotion. Explanations of cars must describe the actual effects of their actual parts. They had best involve, in the case of my car, a combustion engine rather than a platypus, and a fuel injector rather than a carburetor.

What I've shown is that the Epiphenomenal Trap can and should be avoided. The first step is recognizing that the debate over whether or not there can be a possible functionalist explanation of consciousness is misguided. That discourse cannot settle the question of whether mental kinds are functional kinds. The discourse that forms the Trap is based on a mistake about what claims functionalism is entitled to. The result of this is that Functionalists and Anti-Functionalists must argue from a different philosophical stance than they have been for almost forty years.

> *x* functions as a ϕ in *s* (or: the function of *x* in *s* is to ϕ) relative to an analytical account *A* of *s*'s capacity to ψ just in case *x* is capable of ϕ-ing in *s* and *A* appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of *x* to ϕ in *s*. (Cummins 1975: 762)

Amundson and Lauder (1994) follow Neander (1991) in calling Cummins-style functions *causal role functions*.

References

Amundson, R. and G. Lauder. 1994 "Function without purpose: The uses of causal role function in evolutionary biology". *Biology and Philosophy* 9, 443-469.

Brandon, R. N. 1990. *Adaptation and Environment.* Princeton: Princeton University Press.

Churchland, P. M. 1988. *Matter and Consciousness, revised edition.* Cambridge, MA: MIT Press.

Cummins, R. 1975. "Functional analysis". *The Journal of Philosophy* LXXII, 20: 741-765.

Gould, S. J. and R. C. Lewontin. 1978. "The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist program". *Proceedings of the Royal Society, London* 205: 581-598.

Lycan, W. G. 1987. *Consciousness*. Cambridge, MA: MIT Press.

Neander, K. 1991. "Functions as selected effects: The conceptual analyst's defense". *Philosophy of Science* 58: 168-184.

Putnam, H. 1960. "Minds and Machines". In S. Hook (ed.), *Dimensions of Mind*. New York: New York University Press.

Van Gulick, R. N. 1988. "A functionalist plea for self-consciousness". *The Philosophical Review* XCVII, 2: 149-181.