# Naturalistic Philosophy of Mind: Taking Scientific Explanation Seriously

Andrew Richmond

Abstract

Philosophers of mind tend to accept three claims. (1) Views in the philosophy of mind can and must draw support from cognitive science. (2) A central task for the philosophy of mind is to give a metaphysicsof mind: a definition of the mind, or an account of *what it is to be* minded. (3) The most promising approach to that task characterizes the mind as a *computational* and *representational* system. I argue that these claims are only consistent on a naïve view of cognitive science that understands the explanations of cognitive science as metaphysical commitments rather than efforts at causal modeling. Starting from a more serious understanding of cognitive science, I bring out the inconsistence of the three claims and discuss how to move forward by dropping one of them.

## Introduction

Philosophers of mind tend to endorse the following three views:

Naturalism

The philosophy of mind should be *naturalistic*.

Metaphysics

An essential task in philosophy is to give a *metaphysics* of mind.

Computational-Representational Theory of Mind (CRTM)

The most promising approach in the philosophy of mind is *computational* and *representational*.

I will argue that these three views are inconsistent, and that their inconsistency calls for a reconception of the philosophy of mind — one of the views must be abandoned. I’ll also argue that abandoning the third is not a promising way out of the inconsistency, so the required reconception is more radical than anything like anti-computationalism: it is either anti-naturalism, or an abandonment of what most philosophers take to be the central task in the philosophy of mind. I’ll start by clarifying each view in turn. Then I’ll discuss in more detail the nature of cognitive scientific explanation, focusing on representational and computational explanations. That will allow me to draw out the inconsistency in these views, and I’ll conclude by sketching some (necessarily radical) ways to resolve that inconsistency.

## Naturalism

Naturalism comes in different forms, but it usually involves a *pro-scientific* attitude on the part of philosophers. A weak form of naturalism would say that the philosophy of mind should be *consistent* with scientific results — it should not contradict good science. But most naturalists have something more in mind: they think philosophical theories should be *grounded in* or *supported by* scientific work. A commitment to souls, ghouls, or reincarnation could be consistent with scientific work, but it would be too far removed from science to draw support from it. That support needn’t direct— it can be abductive instead: a philosophical theory can draw support from its success unifying or making sense of scientific work — i.e. from *it’s* support of scientific practice. To capture both directions of support, and to keep the notion of naturalism quite weak, I’ll define naturalism as the view that *philosophical theories should be continuous with scientific work*, in that there should be *potential relations of support* between the two domains. In other words, views in philosophy and science should *bear on* each other.

## Metaphysics of Mind

Descartes did not just ask how we could *best* *frame theories* about the mind. Armstrong did not just argue that causal notions were *explanatorily fruitful* in the philosophy of mind. Searle did not just claim that biology helped to *explain* mental capacities, or that computational notions were insufficient to *illuminate* mental processes. In each case, and in the philosophy of mind more broadly, philosophers are interested in understanding what certain mental capacities *are*, and more generally, what the mind *is*. Philosophers want a *metaphysics* of mind; they want to know the *essence* or *nature* of mind. They may ask for *necessary and sufficient conditions* for being minded, a *definition* of the mind, or what *constitutes* mindedness. Regardless of their specific terminology, they want to know what makes minds *minds* — what it is in virtue of which things belong to that category.

Any number of examples could be given, but Jaegwon Kim summarizes this view nicely:

We need a reasonably clear idea about *what mentality is* and what having a thought *consists in.* *What conditions must a creature or system meet* if we are to attribute to it a ‘mind’ or ‘mentality’? … Is there a *general characteristic that distinguishes mental phenomena* from non-mental, or ‘merely’ physical, phenomena? [Kim 2011, 2, emphasis mine]

## Computational-Representational Theory of Mind

For many reasons, philosophers of mind favor a computational and representational approach. They understand the mind in computational terms: algorithms, programs, networks, and so on. And they characterize it as representing the environmental variables those algorithms are defined over, or, similarly, *coding for* features of the environment. I’ll focus on computation in the following, but the same points apply to representation. The point is that, first of all, philosophers understand the mental processes and cognition in general in computational terms, and explain it using computational notions:

Could a machine think? Could the mind itself be a thinking machine? *The computer revolution transformed discussion of these questions*, offering our best prospects yet for machines that emulate reasoning, decision-making, problem solving, perception, linguistic comprehension, and other characteristic mental processes. [Rescorla 2017, 2, emphasis mine]

And so the standard answers to the questions raised a moment ago (What is the mind? What are mental capacities?) are computational. The mind and mental capacities are defined in computational terms. Cognition iscomputation; to be a mind isto be a computational system of some particular type; the nature of mental capacities is the computational structure they instantiate; and so on. To reiterate, we’re not just talking (e.g.) about how best to model the mind; we’re asking about the *nature* of the mind, we’re asking *what it is* *to be* a mind. Examples of this approach abound, but here are two:

The mind is literally a classical computational system – an interpretable, formal, symbol manipulator – of some sort; and cognitive processes, such as reasoning and visual perception, just are classical computational processes of some sort. So construed, [computationalism] is a kind of *empirically motivated*, *metaphysical* doctrine, in that *it provides a general characterization of what it is to be a mind, or cognitive process*. (Samuels, 2019, p. 106)

[Computationalism] is not intended metaphorically. [Computationalism] does not simply hold that the mind is *like* a computing system. [Computationalism] holds that the mind *literally is* a computing system. [Rescorla 2017, 9, emphasis mine]

Two of our target views are already combined in these quotes: Metaphysics and CRTM. What happens when we add Naturalism? We claim that this metaphysics of mind must be supported by cognitive science, or at least continuouswithit. It is essential that this does not simply amount to claim that cognitive science must use computational and representational notions — *it must use them in a way that supports the metaphysical claims that philosophers make using those notions.* There are two ways that this could work.

First, the scientific use of these notions could be straightforwardly metaphysical, meaning that philosophical views of the mind are a fairly innocuous inference: scientists claim that the brain is a computer, and the brain appears to be the instantiator of the mind, so it is plausible that the brain’s computer-hood *is* its mindedness. Second, the scientific use of these notions could be metaphysically neutral, but a good philosophical theory of the cognitive sciences — explaining their successes, their unity, and so on — will understand the mind as a computer. That means the philosophical views under discussion are arrived at by a somewhat more complicated, but nonetheless legitimate, inference to the best explanation: in investigating cognition, the cognitive science use computational notions in a central, ineliminable and successful way, and the best explanation of these facts is that cognition *is* computation.

What absolutely can’t be the case is that the cognitive sciences use computational notions as a way of bringing in certain modeling tools, and that the ubiquity and success of those modeling tools is explicable simply by reference to (1) cognitive science’s goals and (2) features of those modeling tools that make them a particularly efficient way to achieve those goals. If that were the case, the scientific use of computational notions would not license any inference about the nature of the mind, except that it is well and fruitfully modeled using the tools that computational notions introduce. And this fact would not be explained by cognition being essentially a computational process, but by the metaphysics-neutral features of the models computational notions help us build.

If you’re reading this as Chapter 3 of my dissertation, you will be ahead of me here. If you’re reading this as a stand-alone article, I should clue you in. Chapters 1 and 2 of my dissertation argued (quite convincingly, I might add) that computational and representational notions function, in the cognitive sciences, in exactly the way I just described. They introduce modeling techniques that are fruitful and successful because of metaphysics-neutral facts about cognitive science’s goals —goals that the relevant modeling techniques are particularly good at achieving. Understanding the brain in computational terms is fruitful in cognitive science because it provides a method of building perspicuous models that capture the kinds of causal structures cognitive scientists seek to characterize. Resources from computer science and related disciplines provide a language, concepts, and heuristics, that are well-suited to cognitive science’s purposes because they are well-suited to framing algorithmic explanations of systems’ capacities to map certain groups of inputs to certain groups of outputs. And, of course, this use of computational notions does not support a claim that the brain *is* a computer or that the mind or cognition is computational — just that the brain, and cognitive processes, are fruitfully modeled using tools that computational notions bring with them.

Now, am I just reiterating the age-old claim that the computational approach to the brain is *just a metaphor*? No. The point is that a scientific field will take resources and explanatory strategies from a broad range sources, and the fact that cognitive science’s modeling strategies are derived from the computational sciences does not mean their models imply the brain is a computer, any more than models of traffic derived from fluid mechanics imply that traffic is a fluid, or models of disinformation in virological terms imply that disinformation is a virus. These are examples of domain transfer — the use of tools, concepts, and resources developed for one domain in another. This is a sophisticated and legitimate epistemic strategy, and its use does not in any way impugn a field of science or imply that it is relying on metaphors *rather than* real theories or models.[[1]](#footnote-1) Domain transfers are one way a field comes by its explanatory resources.

## Upshots

So a metaphysics of mind cannot be naturalistic andcomputational-representational. That means one of the views in our must be dropped, which means there are three ways forward. I’ll consider each individually.

First way forward: we keep (1) and (2) and drop (3), rejecting computationalism as a metaphysics of mind. What are our options, then, as metaphysicians of mind? Many approaches to the metaphysics of mind will suffer from the same problem: not all (seemingly very few) scientific notions are metaphysically-laden. One thing cognitive science is certainly committed to is the *causal* nature of the mind. Causal explanation is not just a *strategy* in cognitive science; it reflects a commitment to the causal nature of the processes the cognitive sciences model. Perhaps the view that the mind is a dynamical system is a promising option too. The problem is that virtually every system is causal and dynamical. The question will be: what makes minds *unique*? The previous idea was that their computational properties did this. Perhaps their causal structure will?

We would have to define the relevant causal structure carefully. It can’t just be *our* causal structures that define minds — we don’t want to be too provincial. And we would have to select the right level of causal organization — chemical microstructure isn’t relevant, e.g. But it’s problems like these that the computational theory of mind was supposed to solve — a *computational structure* can be implemented in very different causal structures, and automatically identifies a certain level of organization. So we would need some other notion, in common use in the cognitive sciences, that can do this work for us. The other option is the deflationary one: we say that to be minded is *to have a causal structure* *that we can fruitfully model, given our particular goals, using computational notions*. This would make the metaphysics of minds perspectival, since our choices of modeling strategy depend essentially on our explanatory goals and various other human constraints. I don’t think this should be ruled out, but it will be unacceptable to most readers.

In short, the first way forward suffers from a lack of replacements for computational notions, the problem that any replacement would plausibly suffer from the same problem. And there is no escaping out a deflationist back door.

Second way forward: we keep (2) and (3), but reject (1). We accept that the philosophy of mind floats free from scientific inquiry. This is not such an unusual view. Lots of work has assumed this something like this is the case — e.g. views that emphasize the normative features of the mind. This is a perfectly viable option, with a body of work to support it. But it requires a decision about the constraints that a metaphysics of mind should be subject to. Should a theory of mind support psychiatric efforts? Should it make sense of social or normative phenomena? Should we have a different metaphysics of mind for each purpose? Or is there just a single thing that the mind is, picked out by intuition or the common-sense concept of the mind?[[2]](#footnote-2)

Also, note that the computational view seems to have lost its motivation here. It was supported primarily by the way computational notions figured into the cognitive sciences. And that consideration won’t figure in on this view. So this option may amounts to retaining (2) *alone*, with the particular metaphysics of mind determined by the purposes it is meant to serve and the phenomena it is meant to explain, or by some more universal, non-empirical, constraints.

Third way forward: we accept (1) and (3), but drop (2), leaving behind the metaphysical approach to the philosophy of mind. What are the alternatives to the metaphysical approach? Elsewhere, I’ve approached the computational theory of mind in philosophy as a methodological contribution to cognitive science, exploring the epistemic virtues of different approaches to the study of mental capacities. There is an extremely interesting body of work taking this approach (Baker et al., n.d.; Cao, 2019; Chirimuuta, n.d.; Poldrack, 2020) [Hardcastle]. Cognitive science can be subject to intense and revealing philosophical scrutiny even if metaphysics is out of view.

Alternatively, we could understand the philosophy of mind as bringing cognitive scientific work to bear on other questions of enduring philosophical importance [Ismael, Paul, Burnston]. In many cases these questions are about decision-making: what constrains our decision-making [Ismael], how can we do a better job of making decisions [Paul], and what in general can we say about the human capacity to *decide*, as opposed to simply being caused to do one thing rather than another [Burnston]? Cognitive science can give us important tools for thinking about ourselves, and revealing explanations of our mental lives, without giving us a metaphysics of mind.

On these approaches, we leave aside the metaphysics of mind to investigate mental capacities (on the latter approach) or scientific work itself (on the former approach). Note also that the third way forward is consistent with the second: we can have a metaphysics of mind for whatever purposes we like, as long as we do not understanding it as continuous with scientific work; and we can do the philosophy of mind in a way that is continuous with scientific work, as long as we do not take ourselves to be giving a metaphysics of mind.

## Conclusion

We cannot have a naturalistic and computational metaphysics of mind. Very likely, we cannot have a satisfying naturalistic metaphysics of mind at all. But we can have a lot in its place. We can have a deep and philosophically sensitive metaphysics of the mind, responsive to normative, social, legal, and epistemic considerations. And, perhaps more importantly, we can have illuminating explanations of our mental lives as long as we don’t take them to be mere definitional or taxonomic explanations, and we can have rigorous philosophical investigations of cognitive science as long as we don’t take those explanations to be, again, definitional or taxonomic endeavors.

1. See Barwich & … here? [↑](#footnote-ref-1)
2. Compare the metaphysics of race. Freed from the constraints of biology, it is responsive to important sociological considerations rather than fruitless attempts at identification with genetic constructions. [↑](#footnote-ref-2)