Functionalism, Functional Isomorphism, and Multiple Realizability

Joseph Schilz

Functionalism, the idea that mental states are characterized by their function, formally arose in the mid 20th century in response to earlier attempts at demystifying the mind(Levin, 2009). Many felt that behaviorism had failed to describe the mind at an appropriate conceptual level, or that attempts such as identity theory were overly restrictive upon the conditions of mental states(Levin, 2009). Functionalists declared that the particular physical processes underlying mental activity in an entity were inconsequential; what mattered was the functional role a state played in that entity's existence.

Divorcing form from function meant that these functions could be realized on a multitude of physical forms. Called multiple realizability, many considered this concept to be a virtue of functionalism(Bickle, 2006). It satisfied our intuitions that, for example, if we were to encounter an alien, it could have pain even if its physical mental machinery were completely different from our own. Critics, however, began to charge that the entailment went too far. It allowed for "crazy realizations" of minds. This, they claimed, was functionalism's Achilles' heel, a *reductio ad absurdium* of functionalism.

In this paper, I don't try to prove that functionalism is true. Opponents of functionalism will always be able to find harbor in the the ineffable, the empirically unobservable. I only wish to show that functionalism is valid, that it's plausible, that it's potentially consistent with our experience of the world.

Functionalism equates functional states with mental states, where a functional state is generally taken to mean some state that relates sensory input to future mental states and actions(Levin, 2009); if two entities are in the same functional state, then they're in the same mental state.

Underlying the notion of being in the "same functional state" is the concept of functional isomorphism. As Putnam defines it, functional isomorphism is a mapping f from one system to another, such that if state a is followed by state b in the first system, $f(a)$ is followed by $f(b)$ in the second system(Putnam, 1975, p. 292).

We can formalize this notion a bit more by appealing to the mathematical formulation

of isomorphism. Let G and H be sets, and the function f be a one-to-one mapping between them. Let $\star$ be the function from G to G that we wish to preserve. Then f is an isomorphism with respect to $\star$ iff $f[\star(a)] = \star'[f(a)]$ for all a in G[1].

I did sneak one extra term into the above: $\star'$. I said that $\star$ is a function from G to G, so let $\star'$ be its equivalent from H to H. f specifies how to translate states from G to H, but $\star$ and $\star'$ specify those states' functional relationships.

In mathematics, the term '$\star$' would traditionally be "overloaded" so as to include both $\star$ and $\star'$. Whether we were talking about the $\star$ from G to G or the $\star$ from H to H should be clear from context. But as I'm going to demonstrate, errors in interpreting functional isomorphism seem to arise from errors in syntax. In fact, $\star$ and $\star'$ are two separate functions, so I'll continue to make the distinction explicit.

Let me engage you in a common mathematical practice. When I see a novel equation like $f[\star(a)] = \star'[f(a)]$, the first thing I do is break it into smaller pieces. Equations are best understood as compositional, and considering it a piece at a time illuminates the equation's syntax.

Alright, $\star(a)$ is pretty simple. As $\star$ is a function from G to G, $\star(a)$ denotes some element of G. Let's call this element b. Then $f[\star(a)]$ is equivalent to $f[b]$. This makes perfect sense as well; f takes elements of G and translates them into elements of H.

Similarly, $f(a)$ translates a into an element of H. And if $f(a)$ denotes an element in H, then $\star'[f(a)]$ is valid and denotes another element of H. And that's outstanding, because we've just confirmed that both sides of the equation represent an element of H. That's valid syntactically. Since we take the equation to be true, the elements on either side are the same.

Now we understand the syntax of $f[\star(a)] = \star'[f(a)]$, but what does it really *mean*? What it means is that it doesn't matter whether we apply $\star$ to a before we translate it into H or after. The end result is still the same element of H. You'll notice I overloaded $\star$ in my explanation. But I think it's necessary here, because the underlying idea is that

[1]The full mathematical concept of isomorphism is actually a bit richer than this. I've given $\star$ as a one place predicate, but it could just as easily take more arguments. For $\star$ as a two place predicate, we would define isomorphism as $f[\star(a, b)] = \star'[f(a), f(b)]$.

$\star$ and $\star'$ are two faces of the same underlying function: one operates within the domain of G and the other H.

How does this concept of isomorphism apply to mental states? Let's reconcile it with Putnam's definition. Call G the set of functional states of a typical human, and H the set of functional states of some sufficiently complex system: a humanoid robot. $\star$ is the function we might call "follows from" in English. And f, naturally, is the function we would use when someone identifies a human function state and asks for the robot's equivalent.

Take note! Though f is the isomorphism, it's really $\star$ that does all of the heavy lifting. $\star$ and $\star'$ must satisfy a very particular relationship in order for an isomorphism to exist. In fact, $\star$ carries an overwhelming amount of information. For example, in a "typical human" it relates a complex state like setting, hunger, a set of standing associations, perceived company present, and how well done your hamburger is with another complex state like satisfaction, an altered set of associations, and inclination to take another bite.

Though $\star$ does not refer to a life history or the complexities of the brain explicitly, it contains this information implicitly. To make this point clear, we could provide as input to $\star$ the state of having just been asked what the subject's first kiss was like, to compose a haiku, or the subject's thoughts on the nature of consciousness[3]. It would respond as the human would. We could divine all empirical evidence of humanity from the system, with enough queries of the function $\star$. And by extension, $\star'$, is just as informationally dense[4].

Consider what we said earlier about the meaning of isomorphism. Given a human state a, it doesn't matter whether we translate a into its robot equivalent and then find the following robot state, or find the human state that follows a and translate that new state into a robot state. Either way, we get the same state.

[3]Including, of course, such parameters as being under the impression that the subject is sitting in a room, amicable to speech, has no pressing priorities, etc.

[4]I had to add one thing to Putnam's definition in order to make this true. Under Putnam, I could be functionally isomorphic to a rock. Take $a, b \in G$ as the states "wants ice cream," "decides to get ice cream," and respectively, and specify that $\star(a) = b$. Let $x \in H$ be the state "just sits there," and let $f(a) = x, f(b) = x$ and $\star'(x) = x$. Then $f[\star(a)] = \star'[f(a)]$ would be vacuously true.

The missing ingredient? f must be one-to-one. This prohibits the assignment $f(a), f(b) = x$.

Returning to the robot, it seems easy to see the efficacy of functionalism in this environment. The moment I want to order pizza, I'm in some mental state. Let's say it's state α that I'm in. State α kicks off a sequence of mental states. If we could name these states, some of them might be, "recalling the smell of pepperoni," "imagining a bite of Hawaiian pizza, in an attempt to gauge expected satisfaction versus pepperoni pizza," "briefly recalling the taste of the beer I had with my last pizza," and so on. Eventually, this sequence of mental states eventually impacts my motor cortex in such a way that I pick up the phone and dial a pizzeria. Let's call this sequence A, where A is a subset of G.

If I set the robot with state $f(\alpha)$ and provide it with the same sensory input I received, then it will undergo functional states $f(A)$. And according to functionalists, undergoing the many–perhaps thousands or more–functional states composing $f(A)$ constitutes a mental experience identical to that I experienced when I underwent A.

Few critics deny this example *prima facie*. Instead, they have attacked more abstract realizations, in an attempt to show that multiple realizability, and hence, functionalism are false.

Many of these attacks can be reduced to arguments from personal incredulity. Searle, for example, in responding to critics of his Chinese Room argument, seems to acknowledge that an entity empirically identical to a human could be generated by a formal system. All observable evidence for a mental experience may be ignored, though, because it would lack some unobservable quality that Searle believes himself to possess(Shieber, 2004, p. 211-215).

Elliot Sober provides a slightly more formal attack. He claims that structural isomorphism is not a sufficient condition for the possession of an identical mental state[5]. Evoking Block's "China Mind" thought experiment, he imagines that we've assembled a group of people to simulate Sober's brain as he is think, "I want ice cream. Though

[5]I think that Sober is constructing a straw man in utilizing *structural* isomorphism for his argument. This isn't, after all, "structurism" he's trying to defeat. It's the function we care about.

This is an innocent enough substitution in his argument against sufficiency, because structural isomorphism entails functional isomorphism, but it's the crux of his argument against necessity: the argument fails if you consider functional isomorphism.

his brain and the group are structurally isomorphic[6], "I have the belief," he concludes, "while the group (presumably) does not." (Lycan, 1990, p.64)

Sober is decent enough to recognize that his conclusion rests on presumption. But it highlights a very interesting question: does the group want ice cream?

This is where our formalism will pay off. Remember, I'm defending the validity of functionalism. So, I'm going to proceed by taking up the assumption of functionalism and defending it.

So let's assume functionalism and its entailment, that two entities have the same mental state when they are in a functionally isomorphic state[7]. And let's assume also Sober's simulated brain, as a construct functionally isomorphic to his own brain.

Does the group want ice cream? No, *and we shouldn't expect them to.* Sober has committed an error of syntax in formulating the question. "Wants ice cream" isn't a state that belongs to H, so we should not expect H to be in that state. Is the group in state f("wants ice cream")[8]? By hypothesis, yes! And this is all that functionalism claims.

Sober may have meant to get at a different question than the one he posed. He might have meant to ask if this activity, if this translated wanting of ice cream, carries that same *je ne sais quoi* that we experience when we want ice cream. His answer to this question, apparently a 'no', is the presumption that Sober was referring to.

And how could it possibly? What does it mean for the group to be in state f("wants ice cream")? Sober has given us a rather meager simulation. Let's fill it out a bit before we try to answer that question.

If we're going to simulate Sober's brain, then there's nothing to stop us from simulating an entire small town, perfectly functionally isomorphic to an isolated town in the real world. Whereas our actions occur in real space, let's say that this town exists in "isomorphic space." Of course, the physical basis of this simulation would look nothing

[6] And hence functionally isomorphic

[7] Where two states a, z are "functionally isomorphic" if they belong to the sets of states of two entities which possess a functional isomorphism, and $f(a) = z$.

[8] f("wants ice cream") would be a particular kind of f("wants") that is directed at the isomorphic equivalent of ice cream

like a real town to us. It would look like trillions of people scrambling over the globe and hitting each other with hammers[9].

Just so, the inhabitants would report to each other the reality of the town, its sights and sounds. Imagine a small boy, growing up there. He reports to his mother that his favorite day is Sunday, because every Sunday his father takes him out for ice cream. He's often torn on whether to use his allowance to buy his dog a bone, or to buy ice cream, but sometimes he buys ice cream and shares it with his dog. His first date was at the town's ice cream parlor.

So now, if you don't think that the boy's want for ice cream has the same mental character as your own want for ice cream...well...if you think that then you've denied the premise. You've denied the premise of Sober's *reductio ad absurdium*, the premise that functional states are equivalent to mental states. In the end, Sober can argue only from personal incredulity.

We associate so well with the robot and so poorly with Sober's simulation because the robot's actions are occurring in real space. When we set the robot to its equivalent of wanting pizza, we can point to the thing that would satisfy it. The group of simulators is different; they have set up some persistent pattern of hitting each other over the head that is waiting for a particular change in how the brain stem simulators are behaving. It's hard to see what ice cream has to do with any of it.

Viewed from within isomorphic space, however, it's very clear. If we gave Sober a body with which to act, he would tell us exactly what it's like for him to want ice cream. If we could translate ice cream into the space, it would eagerly eat it. Does it experience the ineffable? If we ask it, by hypothesis it would behave just as Sober in real space and answer with a convicted, "Yes!"

It's probably never going to be possible to prove the truth of functionalism. In practice we will never be able to eliminate the possibility of *some* other influence. And even if we could, epiphenomenalism is an impregnable last stand for the theory's opponents. But functionalism is consistent and valid. And if even its critics seem to accept that it

[9]Per Sober's design.

could account for all human behavior[10], there seems to be no cause to provide additional explanations.

References

Levin, Janet (2009 April 09). *Functionalism.* Retrieved September 1, 2009, from Stanford Encyclopedia of Philosophy Web site: http://plato.stanford.edu/entries/functionalism/

Bickle, John (2006, July 27). *Multiple Realizability.* Retrieved September 1, 2009, from Stanford Encyclopedia of Philosophy Web site: http://plato.stanford.edu/entries/multiple-realizability/

Putnam, Hillary (1975). *Mind, Language, and Reality.* Cambridge, United Kingdom: Cambridge University Press.

Shieber, Stuart (Ed.). (2004). *The turing test: Verbal behavior as the hallmark of intelligence.* Cambridge, Massachusetts: MIT Press.

Lycan, W.G. (Ed.). (1990). *Mind and cognition: An anthology* Malden, Massachusetts: Blackwell Publishers.

Dennett, Daniel (1992). *Consciousness explained.* Ney York: Back Bay Books.

[10] "All human behavior" would include all reports of subjective experience (Dennett, 2009).