13

Must Functionalists Be Aristotelians?

Robert C. Koons and Alexander Pruss

Functionalism remains the most promising strategy for 'naturalizing' the mind. We argue that when functions are defined in terms of conditionals, whether indicative, probabilistic, or counterfactual, the resulting version of functionalism is subject to devastating finkish counter-examples. Only functions defined within a powers ontology can provide the right account of normalcy, but the conception of powers must follow classical, Aristotelian lines, since the alternative (an evolutionary account of normativity as proposed by Ruth Garrett Millikan) is inconsistent with a plausible principle of the supervenience of the mind on local conditions.

1. Functionalism

Naturalizing the mind demands that the fundamental vocabulary of psychology must be wholly physical (for description of inputs and outputs), plus the language of causation, dispositions, counterfactuals, or function, as well as the terms of logic and mathematics, achieving as a result a so-called 'topic neutral' language. British philosopher and logician Frank Ramsey (1929) offered the logical tools needed to express mature functionalism, describing a logical process that has come to be known as 'Ramsification.' We start with the true (and at present not fully known) theory of psychology, one including explicitly mental terms and predicates (like 'pain' or 'conscious of'). This theory is supposed to capture the one Pattern of interactions that is definitive of having a mind. We form a single, gigantic conjunction of all of the postulates of the theory and then replace each mental predicate by a second-order variable of the same type (i.e. one-place predicate variable for monadic predicates, two-place predicate variables for binary predicates, etc.). Finally, we append a series of existential quantifiers to the beginning of the formula, one quantifier for each variable-type. The resulting 'Ramsey' sentence is now in a topic-neutral language, since the only predicates that remain are either part of the language of physics and mathematics, or belong to a category of causal or modal language, such as causal predicates, probabilistic connectives, nomological necessity operators, or subjunctive conditional connectives.

Clauses of the Ramsey sentence will have a form something like one of these:

(1) If the system *x* is in internal state S_n and in input state I_m at time *t*, then *x* at the next relevant time *t*+1 is in internal state S_k and output state O_j. (Indicative conditional)

(2) If the system *x* were in internal state S_n and in input state I_m at time *t*, then *x* would at time *t*+1 be in internal state S_k and output state O_j. (Subjunctive conditional)

(3) Whenever the system *x* is in internal state S_n, *x* has a disposition to enter immediately into output state O_j and internal state S_k in response to input state I_m. (Dispositional state)

(4) System *x*'s being in internal state S_n confers upon it the power to produce output state O_j and internal state S_k immediately in response to input state I_m. (Causal power)

Here, *x* is either the whole mind or a subsystem. This could in principle be a very low level subsystem, say a logic gate that takes two truth value inputs and returns their disjunction, or a very high level one, say one that takes desires and beliefs and outputs motor activation signals.

The project of Ramsifying psychology raises a number of questions, including the following:

- What sort of language is involved in the specification of the links between inputs, internal states, and outputs?
- Does the theory make use of material or subjunctive conditionals, or does it make reference to causal powers or intrinsic dispositions?
- If it does make use of causal powers or intrinsic dispositions, how are these to be understood?
  - In a Rylean way, as equivalent to the truth of a subjunctive conditional?
  - In a Dretske–Armstrong–Tooley way, as following from facts about primitive laws of nature?[1]
  - Or in an Aristotelian way, according to which powers are intrinsic properties of substances, definable in terms of their inputs and outputs, and conferred on substances by essential or accidental natures, and dispositions are powers together with a teleological directedness towards their exercise?

First we will argue that the material and subjective conditional views are untenable, and then we will evaluate the dispositional and powers views. We will argue that the

[1] See Dretske (1977), Tooley (1977), and Armstrong (1983).

only plausible form of functionalism requires that the connections between inputs, outputs, and mental states be described as causal powers, in accordance with the assumptions of standard Aristotelian metaphysics.

2. Conditional Functionalisms

2.1. Material Conditionals

Functionalisms built on indicative or subjunctive conditionals have little hope of success. The most obviously unsuccessful are *material* conditional accounts, simply because the conditional clauses will be satisfied by any system that never actually receives the inputs. The moon will count as a human-level mind, just one that never actually gets to think about anything because the activation conditions are never satisfied.

2.2. Non-material Conditionals

Standard problems with conditional accounts of dispositions apply just as well to all the non-material conditional forms of the accounts. We can imagine, for instance, that the individual has strapped to her a bomb that explodes if system *x* is in internal state S_n and receives input I_m at time *t*, but that in fact this condition does not obtain. Then, the subjunctive conditional (2) is false, and (1) will also be false on plausible non-material readings (e.g. ones based on conditional probabilities). Yet, having such a bomb that never goes off strapped to one, while unfortunate, does not make one not have a mind.

One might try to use context-sensitivity to ward off such worries, for instance by saying that in evaluating conditionals or conditional probabilities we should only consider those causal factors that are internal to the system *x*. But we can replace the bomb by a fatal disease, and the distinction between 'internal' and 'external' causal factors will become untenable.

What if the antecedents of the conditionals are strengthened to include the claim that the whole system survives until the next relevant time? Here we borrow an idea from Harry Frankfurt (1969): the introduction of a purely hypothetical neural-manipulator. The manipulator wants the subject to follow a certain script. If the subject were to show signs of being about to deviate from the script, then the manipulator would intervene internally, causing the subject to continue to follow the script. Moreover, if by some near-miracle the subject succeeded in deviating from the script for a step, the manipulator would push the subject right back to the script. We are to imagine that the subject spontaneously follows the script, and as a consequence, the manipulator never intervenes.

Frankfurt introduced such a thought experiment to challenge the idea that freedom of the will requires alternative possibilities. We use it to show that the existence of mental states is independent of the truth of counterfactual conditionals linking the states to inputs, outputs, and each other. It is obvious that the presence of an inactive manipulator cannot deprive the subject of his mental states. However, the manipulator's presence is sufficient to falsify all of the usual non-material conditionals and conditional probabilities linking the states. If the manipulator's script says that at time *t*+1 the subject is to be in state S_n, then that would happen no matter what state the subject were in at time *t*.

Again, it won't do to say that the conditionals need to hold on the assumption of no external interference (see Smith 2007). For we can always replace an external intervener by an internal one—say, an odd disorder of the auditory center of the brain that causes it to monitor the rest of the brain and counterfactually intervene.

Moreover, cognitive malfunctioning is surely possible as a result of injury or illness. The theory to be Ramsified cannot plausibly incorporate the effects of every possible injury or illness, since there are no limits to the complexity of the sort of phenomenon that might constitute an injury or illness. Injury can prevent nearly all behavior—so much so, as to make the remaining behavioral dispositions so non-specific as to fail to distinguish one internal state from another. Consider, for example, locked-in syndrome, as depicted in the movie *The Diving Bell and the Butterfly*. Therefore, the true psychological theory must contain postulates that specify the *normal* connections among states.

Without resorting to Aristotelian or evolutionary teleology (an option we will discuss later), our only account of normalcy will be probabilistic. Thus, a system *normally* enters state S_m from state S_n as a result of input I_m provided it is *likely* to do this. However, serious injury or illness can make a malfunctioning subsystem rarely or never do what it should, yet without challenging the status of the subsystem as, say, a subsystem for visual processing of shapes. And, again, a merely counterfactual intervener, whether external or internal, can change what the system is likely to do without manipulating the system in any way.

Alternately, one might try to define normalcy in terms of what systems *of the same type* are likely to do. Thus, a system *normally* enters state S_m from state S_n as a result of input I_m provided that most of the time systems of *this type* do this. A serious problem here is that we are giving the functional claims in order to characterize the *type* of system. But it is then circular in the functional claims to refer to other systems of the same type. One might try to Ramsify over types to solve this problem, but one will still have problems with one of a kind minds.

Moreover, the probabilities of state transitions in systems of a given kind depend deeply on the environment the systems are in. A plausible account would have to say that a normal transition is one that is likely to occur in systems of the given type *in a normal environment*. But, again, it does not appear possible to specify a normal environment without resorting to something like teleology or proper function.

2.3. Rylean Conception of Dispositions

Rylean dispositions (see Ryle 1949) correspond to the subjunctive conditional: if *C* were to be realized, then *E* would result. Hence, Rylean dispositions are also subject to the objections to conditional views when used to formulate functionalism.

2.4. Nomological-deductive Model of Powers

A thing or system of things *S* has the *C*-to-*E* nomological-deductive disposition if and only if there is some description *D*(*S*) satisfied by *S* and laws of nature *L* such that *L*&*C*&*D*(*S*) entails *E* (or, perhaps, such that the rational probability of *E* on *L*&*C*&*D*(*S*) is constrained to be very high). Again, the bomb and fatal disease objections to conditional views rule out nomological-deductive dispositions when these are used to formulate functionalism.

3. Three Theories of Normativity

There are three plausible accounts of the basis of normativity: Aristotelian powers, agential intentions, and evolutionary accounts.

3.1. Aristotelian Normativity

An Aristotelian can give a straightforward account of normativity: a substance is supposed to produce *E* on occasions of *C* if and only if its nature includes a *C*-to-*E* power (one might also prefer more active terms like ‘tendency’ or ‘striving’).

This account may appear insufficient in the light of the possibility of indeterministic powers. Could not a substance have both a *C*-to-*E* and a *C*-to-non-*E* power, in which case it would neither be supposed to produce *E* in *C* nor to produce non-*E* in *C*? One might complicate the account by excluding such cases of competition in some way, or positing higher order powers that decide between the competing powers. But there are also two simpler moves. One move is to say that in such cases, the substance is in the ‘unhappy’ position of being supposed to do incompatible things—it will necessarily fail at one of them.

A more complex move is to say that it cannot happen that a substance has both a *C*-to-*E* and a *C*-to-non-*E* power. Rather, the substance has a *C*-to-*E* and a C'-to-non-*E* power, and if it happens that both *C* and C' obtain, then the substance will fail to do one of the things it should do. This move fits with a natural metaphysical interpretation of quantum indeterminacy. Take an electron in the mixed spin state |up>+|down>, and measure the electron’s spin, thereby forcing the electron’s state to collapse indeterministically to |up> or to |down>. Suppose the electron ends up going to |up>. What explains its going to |up> is not that the electron used to be in state |up>+|down>. Rather, what explains its going to |up> is that the electron used to be in a state that had an |up> component (or had a significant such component). That the state also had a |down> component is true but does not help to explain the electron’s transitioning to |up>. Thus, the electron has two powers with incompatible outcomes and different, but potentially co-occurring, activating conditions: (a) being in a measurement situation with a state with an |up> component and (b) being in a measurement situation with a state with a |down> component.

Functionalism can then be put in an Aristotelian mode, referring to the presence of powers to produce outputs and internal states (including other powers). The result would be a non-reductive and non-physicalist version of functionalism, since the form of the theory would rule out the states' realizers being merely physical states of constituent particles (see Bealer 2010).

3.2. Agential Normativity

Normativity of a kind arises from agents' intentionally making and using things:

(5) A thing is supposed to produce *E* on occasion *C* if and only if its maker or users intend it so to do.

For example, hammers are supposed to drive in nails, since this is what the makers and users of hammers intend to do with them. There are two problems with incorporating this kind of normativity into our universal psychological theory of the Pattern of mind. First, it would make it a matter of metaphysical necessity that every mindful thing is an artifact, made and intended to be used by other agents. Second, it would generate an infinite regress, since the agents who are using the mindful things must themselves have minds, necessitating that they too are artifacts made and used by still earlier agents. The regress (or circularity) is vicious, since the relevant norms never acquire any content.

A functionalist might instead try to make use of Wittgensteinian norms which arise from communal rather than individual agency:

(6) A thing *x* is supposed to produce *E* on occasions *C* if and only if there is a game *G* in which *x* is a participant in role *R*, and *G* includes the rule that participants playing role *R* produce *E* on occasions *C*.

Presumably, a game's including such a rule consists in its participants' believing that others will satisfy the rule, and intending to satisfy it themselves, conditional on its satisfaction by others. (See David Lewis's [1969] *Convention*.) This again results in a vicious regress or circularity if all mental activity is supposed to be dependent on the presence of such normativity. In other words, just as we saw for individual agential normativity, while there can be cases of this sort of normativity, it cannot be that this normativity is foundational with respect to the mental life.

Furthermore, surely some solitary animals, such as sharks, have mental properties, even though they do not participate in any Wittgensteinian games.

3.3. Objections to Evolutionary Accounts of Normativity

The third and final potential source of normativity is evolutionary selection. If a system *x* belongs to a reproductive family *F*, then *x* is supposed to produce *E* under circumstances *C* if and only if doing so is one of *F*'s adaptations. This seems to be the most promising alternative to the Aristotelian account, since there doesn't seem to be any vicious circularity or regress.

Ruth Garrett Millikan developed such an account in considerable detail (in *Language, Thought, and Other Biological Categories* [1984]). Here is a simplified version of her definition (1984, 28), which will be a paradigm of such accounts of normativity:

(7) A thing *x* is supposed to produce *E* in circumstances *I* if and only if (i) *x* belongs to a reproductive family *R* in which some feature *C* occurs non-accidentally with nontrivial frequency (i.e. strictly between 0 and 1), (ii) there has been a positive correlation between having feature *C* in *R* and producing *E* in circumstances *I*, and (iii) this positive correlation has been in part causally responsible for the successful survival and proliferation of family *R* (including *x* itself).[2]

Similar proposals have been made by Larry Wright (1973), Karen Neander (1991, 1995), Nicholas Agar (1993), Kim Sterelny (1990), David Papineau (1993), and Fred Dretske (1995). Here, for example, is Neander's definition:

(8) Some effect (*Z*) is the proper function of some trait (*X*) in organism (*O*) iff the genotype responsible for *X* was selected for doing *Z* because *Z* was adaptive for *O*'s ancestors. (Neander 1995, 111)

Neander distinguishes a range of options for the evolutionary account of function, from what she calls the "High Church" approach of Millikan to her own "Low Church" version (1995, 126–36). The two versions differ by restricting the genuine proper functions to those corresponding to the 'highest' level description meeting definitions (7) or (8) (the "High Church" option) or to the 'lowest' level (the "Low Church" option). Higher-level descriptions refer to more remote effects, such as being able to find suitable nutrition, while lower-level descriptions refer to more proximate effects, such as accurately indicating the presence of an opaque moving body. Our objections apply to both versions as well as to the "Broad Church" option, which would count all levels as containing genuine proper functions.

There is a further distinction between historical or backward-looking accounts (including all of those mentioned above) and the forward-looking account of Bigelow and Pargetter (1987). Forward-looking versions of (7) and (8) are easy to generate: simply replace the past-tense references to causal contributions to the survival and

[2] Millikan's actual definition requires that *C* be a 'Normal' or reproductively established characteristic of *R*. Instead of requiring that *C* be positively correlated in *R* with the function *F*, she requires only that the positive correlation hold in some set *S* which includes *x*'s ancestors, together with "other things not having *C*." Her exact wording of clause (3) is:

> One among the legitimate explanations that can be given of the fact that [*x*] exists makes reference to the fact that *C* correlated positively with *F* [i.e. the function of producing *E* in circumstances *I*] over *S*, either directly causing reproduction of [*x*] or explaining why *R* was proliferated and hence why [*x*] exists. (Millikan 1984, 28)

None of these variations would make any difference to our objection.

reproduction of ancestors with present-tense references to an increased propensity to survive and reproduce on the part of existing members of the population. Most of our objections will apply with equal force to the forward-looking version. And there is a special objection to forward-looking accounts, based on the following dilemma: either the dispositions to reproduce are defined in relation to the 'normal environment' of the species, or not. If so, the account is viciously circular, since an environment is normal for the species only if members of the species are disposed to reproduce in it. Alternatively, if we define the forward-looking dispositions in relation to the organisms' actual environment (whether normal or not), then we get the absurdity that we can tell a priori—simply by observing that we still have a mental life—that our external environment is still normal.

There are a number of objections to these evolutionary accounts.

Objection 1: Can 'reproduction' be defined naturalistically and without reference to function or teleology? Complex organisms (especially ones that reproduce sexually) never produce exact physical duplicates of themselves. Conversely, since everything is similar to everything else in some respects, every cause could be said to be 'reproducing' itself in each of its effects. Real reproduction involves the successful copying of the essential features of a thing. For living organisms, these essential features consist almost entirely of biological functions. Hence, we cannot identify cases of biological reproduction without first being able to identify the biological functions of things. Yet, Millikan's account requires us to put the reproductive cart before the functional horse.

A Millikanian version of functionalism would have the consequence that a thing has a mind only if it belongs to a reproductive family *R* for which the standard Pattern of dispositions has successfully contributed to the survival of *R*. Thus, whether a thing has a mind depends on the evolutionary history of its kind. This engenders a second problem.

Objection 2: Millikanian functionalism (i.e. the backward-looking version of the evolutionary account) has the implausible consequence that mental functioning is one generation behind neural functioning. For a mutation can never be normal on her account in the generation in which it first occurs—it only becomes normal in their descendants. For instance, on this view, presumably one of our distant vertebrate ancestors, call it Sim, evolved the first form of those neural structures that are responsible for consciousness. But it was Sim's children, not Sim, that were conscious if we use Millikanian functions as the backing for functionalism. For on Millikanian views, the structures as found in Sim did not function normally. It was only once their non-normal functioning helped Sim reproduce that they functioned normally in Sim's descendants and hence made them conscious. Not only is this an implausible claim, but it has an undesirable epiphenomalist consequence. Consciousness as such is useless to us—it does not affect our action or fitness. Assuming Sim's children had no relevant new mutations, their behavior was much like Sim's, but they were conscious while Sim was not.

Objection 3: What does it mean for a particular disposition to 'cause' or to 'contribute' to a particular instance of R-reproduction? There are two possible answers. First, we could say that the disposition contributed to the act of reproduction just in case some exercise of the disposition by the parent occurs in the actual causal history of the creation of the child. Second, we could instead require that the disposition be part of a *contrastive* explanation of the reproduction: part of a minimal explanation of why in this instance reproduction or survival occurred, as opposed to not occurring. (The forward-looking version of Bigelow and Pargetter must rely on contrastive explanations, since that is the only way for a trait to contribute to the present propensity to survive and reproduce.)

The first answer would greatly over-generate adaptations. Any feature of the parent that is both the product of some disposition of the parent and that influences in any way the process of reproduction would count as one of the kind's essential adaptations. For example, suppose that rabbits are disposed to twitch their left rear leg whenever a cosmic ray strikes the spinal cord at a single point, and suppose that this disposition was actually exercised by some rabbit in the past as it was successfully locating a bunch of carrots. Even if the twitch played no role in explaining the rabbit's survival, it would still count as adaptive, so long as it was part of the total cause of this rabbit's survival in this concrete instance.

Thus, we'll need to turn to the second answer, contrastive explanations. The use of contrastive explanation fits standard biological practice, which identifies adaptations with the results of natural selection, and selection is inherently contrastive in nature.

Now to our objection. Say that a region R of spacetime is *impotent* provided that nothing in R can affect what happens in spacetime outside R. Consider first the following principle:

(9) (Almost global supervenience of physical minds.) Suppose worlds w_1 and w_2 are exact physical duplicates, except in an impotent region R of spacetime. Then w_1 contains an instance of mindedness outside of R if and only if w_2 contains an exactly similar instance outside of R.

Imagine a world w_1 which contains a planet much like earth, where history looks pretty much like it looks on earth, and which also contains a Great Grazing Ground (GGG), which is an infinite (we only need: potentially infinite) impotent region. Moreover, by a strange law of nature, or maybe by the activity of some swamp aliens, whenever an organism on earth is about to die, it gets hyperspatially and instantaneously transported to the GGG, and a fake corpse, which is an exact duplicate of what its real corpse would have been, gets instantaneously put in its place on earth. (We will call it 'earth' for convenience but we shan't worry about its numerical identity with our world's earth.) Furthermore, there is no life or intelligence outside of earth and the GGG.[3] Moreover, the organism dies as soon as it arrives in the GGG.

[3] Assume that any swamp aliens who created the GGG and the transport system don't count as alive or intelligent.

Our world's earth has organisms with real minds, and the earth in w_1 has a history that is just about the same. The only difference is that in w_1 all the deaths of organisms occur not on earth but in the GGG, because they get transported there before death. But this does not affect any selective facts. Thus, the evolutionary theorist of normativity should say that the situation in w_1's earth is similar enough to that on our earth that we should say that w_1's earth contains organisms with exactly the same minds.

The hard work is now done. For imagine a world that is exactly like w_1 outside of the GGG, but inside the GGG, immortal and ever-reproducing aliens rescue each organism on arrival, fixing it so it doesn't die, and even make the organism capable of reproduction again. Furthermore, they do the same for the organism's descendants in the GGG. The GGG is a place of infinite (at least potentially) resources, with everybody having immortality and reproduction, with the aliens shifting organisms further and further out to ensure their survival.

Now in w_2, there is no selection: Nobody ever dies or ceases to reproduce. Thus, by Millikan's definition (7) or Neander's definition (8), on the contrastive reading, there is no mindedness outside the GGG in w_2—all the earthly critters are functionless zombies. But, by principle (9), there must be instances of mindedness outside the GGG in w_2, because w_2 is an exact duplicate of w_1 outside of the GGG. Hence we have absurdity. This same result obtains in the case of the forward-looking definition: since every member of every population has a perfect propensity to survive and reproduce, no specific trait contributes causally to that propensity.

Suppose our evolutionary theorist of mind denies (9). Then we have the following absurdity: It is up to the aliens in the GGG to determine whether or not there are instances of teleology (including cases of mindedness) outside the GGG, by deciding whether to rescue the almost dead organisms that pop into the GGG. But how can beings in an impotent region bring about that there are, or are not, minds outside that region? That would be worse than magic (magic is presumably causal).

In the GGG story with post-transportation rescue, there is no natural selection, but surely there is mindedness. This shows that not only are Millikan-type stories insufficient for functionalist purposes, but *no* story on which the normativity of mental functioning is grounded in natural selection facts has a chance of succeeding.

4. Conclusions

Functionalism is the naturalist's best hope for a theory of mind. However, functionalist accounts of mind cannot merely make mind depend on the actual behavior of neural systems—they need to be based on the *normal* or *proper* behavior of neural systems. And only broadly Aristotelian theories are able to give an account of this normal behavior. The theory we specifically have offered makes use of the teleological concept of a power to *E* in *C*. We might also have considered a view focused on the disposition to *E* in *C*, where this disposition is irreducible, but powers have some additional metaphysical benefits.[4] We could also have opted for a theory that leaves normalcy un-analyzed.

The net result is that the only kind of naturalist theory of mind that is defensible is an Aristotelian naturalist theory of mind. Most contemporary naturalists do not consider Aristotelian naturalism to be a species of naturalism. But perhaps they will reassess this judgment if Aristotelian naturalism is the only hope for a naturalist theory of mind.

[4] For example, Pruss (2011) and other articles in this volume.