COMMENTS AND CRITICISM

FUNCTION ATTRIBUTIONS DEPEND ON THE EXPLANATORY CONTEXT: A REPLY TO NEANDER AND ROSENBERG'S REPLY TO NANAY*

In "A Modal Theory of Function," I gave an argument against all existing theories of function and outlined a new theory.[1] The argument, briefly, is that all existing theories of function define the function of a token trait in terms of some properties of some other (past, present, future) traits of the same type. They define the function of my heart in terms of some properties of other hearts (say, the hearts of my ancestors). Hence, these theories need to give an unproblematic and noncircular account of trait-type individuation: of what makes hearts hearts. But, I argue, the least problematic accounts of trait-type individuation appeal to the traits' function—thus, we cannot use them as part of the definition of function without running into circularity.[2] My positive account is that we need to define the function of a token trait with the help of modal claims—in terms of what this very token trait would do if things were somewhat different. In short, my suggestion is that we should define function in modal terms.

Karen Neander and Alex Rosenberg argue against both my negative and my positive claims.[3] My aim here is not merely to defend my account from their objections, but to (a) very briefly point out that the new account of etiological function they propose in response to my criticism cannot avoid the circularity worry and, more importantly, to (b) highlight, and attempt to make precise, an important feature of my modal theory that may have been understated in the

* This work was supported by the EU FP7 CIG grant PCIG09-GA-2011-293818 and the FWO Odysseus grant G.0020.12N. The title is a pastiche of Karen Neander's "Explaining Complex Adaptations: A Reply to Sober's 'Reply to Neander'," *British Journal for the Philosophy of Science*, XLVI, 4 (December 1995): 583–87.

[1] Bence Nanay, "A Modal Theory of Function," this JOURNAL, CVII, 8 (August 2010): 412–31.

[2] See also Nanay, "Symmetry between the intentionality of minds and machines? The biological plausibility of Dennett's position," *Minds and Machines*, XVI, 1 (February 2006): 57–71; and Nanay, "Function, Modality, Mental Content: A Response to Kiritani," *Journal of Mind and Behavior*, XXXII, 2 (Spring 2011): 84–87.

[3] Karen Neander and Alex Rosenberg, "Solving the Circularity Problem for Functions: A Response to Nanay," this JOURNAL, CIX, 10 (October 2012): 613–22.

0022-362X/12/0910/623–27

original paper—that function attributions depend on the explanatory project at hand.

Neander and Rosenberg argue that if we formulate the theory of etiological function properly, then the trait-type individuation objection could be avoided. The general gist is as follows. Trait tokens form lineages: the lineage to which my heart belongs, for example, also includes my mother's heart, my daughter's heart, and so on. These lineages can be parsed according to the selection pressures operating on them: if there is selection for doing *F* at a certain part of a lineage, then the traits in that part of the lineage have the function to do *F*. And if there is selection for doing *F* at a certain part of a lineage, then the traits in that part of the lineage all count as belonging to the same trait type. In short, there is no circularity here: both function and trait-type individuation "co-supervene" on the selection processes operating on the lineage.[4]

The problem with this argument is that lineages themselves could not be identified without talking about trait types. Take, as an example, my heart. It is on the same lineage as my mother's heart. But when we say so, we already identify an organ in my mother's body as a heart: as a token of the type that my heart is also a token of. Hence, Neander and Rosenberg have a problem; if the identification of the lineage presupposes an appeal to trait types, then the circularity worry has not gone away. If function supervenes on the parsing of lineages according to selection pressure and lineages themselves are identified by means of trait types, then the definition of function presupposes trait-type individuation. We cannot use functional criteria for individuating trait types without running into circularity. The circularity worry applies to Neander and Rosenberg's new account of function as much as it does to the original version of the etiological theory.

If we want to avoid this circularity, we need to define the function of a token trait in terms of the properties of this very token trait only. And I see no way of doing so other than by endorsing the modal theory of function. This theory may have some unusual features, but the alternative, as far as I can see, is to dispose of the concept of function altogether because of circularity.

Neander and Rosenberg also argue against my positive account, the modal theory of function. According to the modal theory of function, the function of a trait depends not on the history of traits of the same kind, but on the modal properties of this very

[4] *Ibid.*, p. 617.

trait. The only way of avoiding the trait-type individuation objection is to define the function of a token trait with reference to the properties of this token trait only—without talking about other traits of the same type. But if we want to (and we do need to) allow for the possibility of malfunctioning—for the possibility that this token trait has the function to do *F* but fails to do *F* at the moment—we need to appeal not only to this token trait's actual properties but also to its modal properties when defining its function.

The definition of function I gave is the following:

> *Performing* F *is a function of organism* O*'s trait* x *at time* t *if and only if some 'relatively close' possible worlds where* x *is doing* F *at* t *and this contributes to* O*'s inclusive fitness are closer to the actual world than any of those possible worlds where* x *is doing* F *at* t *but this does not contribute to* O*'s inclusive fitness.*[5]

I left the notion of "relatively close possible world" explicitly open. I pointed out that depending on the explanatory project at hand, we need to consider different sets of "relatively close possible worlds." What counts as "relatively close" in one explanatory project will not count as "relatively close" in another. As function attribution depends on the explanatory project at hand, a claim Neander and Rosenberg explicitly agree with,[6] this flexibility of considering different sets of possible worlds to be "relatively close" accounts for this variation of function attribution in different explanatory projects.

Neander and Rosenberg argue that my account gives the wrong prediction about some function attributions. Their example is the function of the mechanism that is responsible for lactose intolerance in a lactose-intolerant individual. They say that I am forced to conclude that this mechanism has the function to digest lactose but is malfunctioning. And this is the wrong conclusion to draw because "it is normal for [this individual], given his ethnicity, to have adult lactose intolerance."[7]

The case, they argue, generalizes: take a trait that can do *F* at the moment but at some time in the future will be able to do *F**, which is much more fitness enhancing than *F*. They claim that I am forced to conclude that the function of this trait is *F** already now because doing *F** would contribute to the organism's inclusive fitness. That is, there is a "relatively close" possible world where this trait is doing *F** and this contributes to the organism's inclusive fitness, and this possible world is closer to the actual one than any possible world

[5] Nanay, "A Modal Theory of Function," p. 422.
[6] Neander and Rosenberg, *op. cit.*, p. 614. See also their footnote 3.
[7] *Ibid.*, p. 616.

where this trait is doing *F** and this does not contribute to the organism's inclusive fitness. But this is wrong: the trait does not now have the function to do *F**.

What Neander and Rosenberg seem to overlook in this argument is that the set of possible worlds we need to consider when determining the function of a trait depends on the explanatory project at hand. Suppose that *x* has an intrinsic property (or set of intrinsic properties) *I* and operates in environment *E*. It has the function to do *F*. Now, suddenly (as a result of a mutation), the intrinsic property of *x* changes to *I**. If *x* were to live in an environment *E**, it could do *F**, which is much more fitness enhancing than *F*. Question: what is the function of *x* now?

The answer is that this depends on the explanatory project at hand. If we are interested in the intrinsic properties of *x*, say, the differences between *I* and *I**, we should keep the environmental factors fixed when considering "relatively close possible worlds." Those possible worlds where the environment is different from the actual one are not considered to be "relatively close" for the purposes of this explanatory project. Thus, possible worlds with *E** are excluded from the set of "relatively close possible worlds." But then the function of *x* will be to do *F*—it is not the case that there is a "relatively close" possible world where *x* does *F** and this contributes to the inclusive fitness of the organism. (It is also not the case that this world is closer than any possible world where *x* does *F** and this does not contribute to the inclusive fitness of the organism.)

But if we are interested in the various things the new intrinsic property, *I**, allows *x* to do, then we should keep the intrinsic properties of *x* fixed when considering the set of possible worlds for the purposes of determining *x*'s function. Those possible worlds where *x* has *I* will not count as "relatively close" in this explanatory project. But possible worlds where the environment is *E** could count. Thus, in this explanatory project, *x*'s function will be to do *F**. There is a "relatively close" possible world where *x* does *F** and this contributes to the inclusive fitness of the organism, and that world is closer than any possible world where *x* does *F** and this does not contribute to the inclusive fitness of the organism. The variability of the scope of "relatively close possible worlds" in my definition is not a bug—it is a feature.[8]

Note that the same argument applies if we swap the intrinsic and the environmental factors: if *x*'s environment changes suddenly

[8] I made the same general point in my "A Modal Theory of Function" on pp. 425–26—and Neander and Rosenberg seem to be in agreement with that argument.

to E^* and, while x has the intrinsic property I at the moment, if it were to have I^*, it could do the fitness-enhancing F^*. If we keep the environment fixed (say we are interested in how this new environment could influence the organism's fitness), then x does have the function to do F^*. If we keep x's intrinsic properties fixed (say we are interested in the anatomy of x in a range of different environments), then x will not have the function to do F^*—there will be no "relatively close" possible world where x does F^*. It seems that Neander and Rosenberg only consider explanatory projects of the latter kind when they argue against my account. But in the case of these explanatory projects, the modal theory yields the very same conclusion that they find intuitive: x does not have the function to do F^*. But there are some explanatory projects of the former kind, where we hold fixed the environmental and not the intrinsic features of x. In these cases, the modal theory does entail that x has the function to do F^*. Crucially, my account allows for this variation in function attribution, depending on the explanatory project at hand—something of which Neander and Rosenberg explicitly approve.

BENCE NANAY

University of Antwerp and
University of Cambridge