Penultimate Drafts: Please consult the published version for corrections and for citation purposes:
Garson, J. 2022. Précis of Karen Neander's *A Mark of the Mental*, *Philosophy and Phenomenological Research* 105: 461-467. http://dx.doi.org/10.1111/phpr.12933
Garson, J. 2022. Response to Neander's Critics, *Philosophy and Phenomenological Research* 105: 490-503. http://dx.doi.org/10.1111/phpr.12931

Précis of Karen Neander's *A Mark of the Mental*

In writing a précis and defense of Karen Neander's book, I want to be clear at the outset that I do not purport to speak on behalf of Neander. I don't know exactly how she would have written a précis of her book, or how she might have responded to her critics. Rather, I write this précis and defense from the standpoint of a long-time admirer of her work, as someone who largely endorses the viewpoint about teleosemantics, and the underlying view of biological function, that she defends, and as a colleague and friend who often discussed her work with her. It is possible that I have misunderstood some elements of her book. It's also possible that she would not have entirely endorsed all of the points that I make in defense of her book. This represents my own interpretation and defense of the project she was engaged in.

A Mark of the Mental: In Defense of Informational Teleosemantics, is Karen Neander's long-awaited first book. It represents decades of careful thinking about the nature of mental representation, which she began writing about in a serious way in the mid-1990s (Neander 1995; also see her 2006; 2013), but the roots of which can be found in her 1983 dissertation on philosophical problems of psychiatry. In particular, the book is a presentation and defense of informational teleosemantics (IT hereafter), a theory about the content of sensory-perceptual, or more broadly, "nonconceptual" representations. She does not present a theory of *conceptual* representation, though she broaches the issue near the end of the book. However, she's quite explicit that IT is intended to be a foundation for any such theory.

Mainstream teleosemantics joins two claims. The first is that mental representation is somehow grounded in biological functions. The second is that biological functions are selected effects, usually construed to be evolutionary selected effects. Teleosemantics is *naturalistic*, in that it doesn't appeal to entities outside of those countenanced by natural science, and it's *reductionistic*, in the sense that it seeks to provide a complete or exhaustive characterization of intentionality exclusively in terms of non-intentional properties.

How, then, does IT differ from other versions of teleosemantics, such as those espoused by Ruth Millikan, David Papineau, or Nicholas Shea? The chief difference is that Neander adopts an *input-centered*, rather than an *output-centered*, view of content. To illustrate the difference, consider the neural circuitry that allegedly has the task of detecting snakes in our environments (Isbell 2009). Here, to simplify greatly, we have a neural mechanism that has the function of causing us to enter a certain brain state, and this brain state is *about* something; it has a meaning or content; it means something like, *there's a snake*. The philosophical question here is, by virtue of *what* is this particular brain state about snakes? Output-oriented or "consumer" teleosemantics holds, very roughly, that the brain state is about snakes because the brain systems that utilize that representation (those neural systems "downstream" from the representation, such as the motor system among others) have the biological function of producing snake-appropriate thoughts, feelings, and actions, such as jumping away in fear. Input-oriented, "producer," or informational teleosemantics begins at the input end. It holds (again, roughly) that the brain state is about snakes because there's a neural mechanism that has the biological function of producing that brain state when snakes are present.

Importantly, informational and consumer teleosemantics do not disagree *at all* about the empirical facts about the brain or about how natural selection works. Rather, they disagree about *which of those empirical facts grounds content*. But how could these two views ever come apart? That is, how could informational and consumer teleosemantics ever diverge in their contents, given that they're describing different aspects of one and the same mechanism and the same historical selection process? Here's one scenario (this is a version of a thought experiment found in Pietroski 1992): imagine an Arctic hare that has a random gene mutation that makes it attracted to the green, needle-shaped leaves of spruce trees. Imagine, furthermore, that this mutation spreads throughout the population of hares because, when hares are under spruce trees, they're hidden from an aerial predator. (Let's also suppose that the hare has no special ability to detect this aerial predator until it's actually getting eaten by one.) Now, consider the brain state that's triggered by spruce trees. What's the content of that brain state? What's it about? Input-centered teleosemantics says the content is something like: *there are those green needle-shaped leaves* (since that's what the hare's perceptual mechanisms are sensitive to). Output-based teleosemantics says the content is something like, *there's a predator-free spot* (since that's the effect that explains the mechanism's evolutionary persistence).

There are two reasons that input- centered teleosemantics gives the more satisfying verdict. The first is that it seems to be intuitively correct (*there are those green needle-shaped leaves*). After all, if the hare (by stipulation) has no way of detecting aerial predators, how can it have representations about them? The second, and more important for Neander, is that it yields contents that hang together well with our understanding of cognitive science (construed broadly to include neuroscience, cognitive neuroscience, and computational cognitive science). As a rule, cognitive scientists try to explain how we can represent *deep* features of the world, such as natural kinds or chemical composition, in terms of our more basic ability to represent *surface* features of the world, like lines, edges, movement, color, and shape.

This is enough by way of preliminaries. Now for a chapter-by-chapter overview. In the first chapter, Neander positions her project within a broader, naturalistic tradition for thinking about representation. She outlines a number of quite substantive assumptions that frame her project. For example, she assumes that a good theory of content should be naturalistic, it should be reductionistic, and it should hang together with our best current account of cognition. Although she doesn't emphasize the point here, she also assumes that a good theory of content should be evolutionarily plausible and that it should help us see how our ability to make and use representations could have evolved from simpler, (quasi-)representational, abilities. Neander's attitude toward these substantive assumptions appears to be something like this: instead of trying to justify all of these principles in an *a priori* manner, let's adopt them as working assumptions, and see whether, on their basis, we can construct a coherent and plausible theory of content. If so, that would go some way toward justifying those assumptions. Put differently, anyone looking for an extensive, *a priori* defense of the claim that naturalism is true, or that evolution and cognitive science are the right starting points for approaching the human mind, will have to look elsewhere.

This chapter also introduces some core terminology. Most important, a *representation*, or a *representational vehicle*, is the entity which represents the world. For humans, it is paradigmatically a brain state (e.g., a signature pattern of neural activation). The *content* of that representational vehicle is whatever that representation is about. I'll usually characterize a representation's content as a proposition: *there's that lush green foliage again*. These representations may or may not have any phenomenal qualities. For Neander, intentionality should be understood, at the nonconceptual level, independently of phenomenal qualities. This is a particularly important working assumption, if we wish our theory to help us understand the content of the sub-personal representations that cognitive scientists routinely posit, not to mention the representational powers of creatures that, quite possibly, do not have phenomenal consciousness, such as toads and frogs.

She also clarifies that her theory is a theory of "original intentionality," the kind of intentionality that doesn't depend on any deeper layers of intentionality, rather than "derived intentionality," the kind that does. One reason that Neander does not emphasize the intentionality of *concepts* is because she thinks many concepts inherit their intentionality from the intentionality of sensory-perceptual representations, and therefore the intentionality of concepts is typically of the derived sort and hence outside of the book's scope. The goal here is to expose the bedrock level of intentionality and to show how higher levels may be built upon that foundation.

Chapter 2 aims to debunk a common misconception about the notion of representation in cognitive science. One idea is that the representation-talk that's so prominent in cognitive science isn't about *real* intentionality. At most, cognitive science trades in some purely causal or correlational notion of "information." A key difference between the kind of representation-talk that cognitive science trades in, and genuine intentionality (in this view), is that genuine intentionality is *normative*: it carries within it the possibility of *mis*-representation, or representational error. Neander takes quite a bit of time to establish, by reference to recent empirical work, that at least *some* representations postulated by cognitive science are normative (error-permitting). In particular, she describes a rare visual defect in which subjects systematically misrepresent the position of visual objects. Such examples show that cognitive scientists who study perception often avert to a rich, normative notion of representation, *in addition to* a rather thin, causal or correlational, notion of information.

Teleosemantics seeks to understand representation in terms of biological functions. So, what are biological functions? Chapter 3 is one of two chapters devoted to this interesting topic. This chapter represents a quite substantial shift in focus (though not in doctrine) from her previous work. One would expect, at this point, that she would defend her flagship view, the selected effects theory of function, which holds, roughly, that a trait's function is whatever it was historically selected for. Interestingly, she *doesn't* – or not to any great extent. Rather, she invokes the more skeletal and abstract idea of a *normal-proper* function, and notes that these normal-proper functions are prevalent in biology. Normal-proper functions, whatever else they are, support a function/accident distinction and a function/malfunction distinction. That is, we can distinguish a trait's proper function from its accidental side effects (say, pumping blood and making beating sounds for the heart), and we can talk meaningfully about malfunction or dysfunction. *Admitting of those two distinctions, as it were, exhausts the content of the concept of* *a normal-proper function* (52). The selected effects theory is one theory about what these normal-proper functions are. Teleosemantics, in her view, should now be understood as the claim that representations are grounded in the normal-proper functions of the brain. Of course, she's still committed to the view that some version of the selected effects theory is the best theory of normal-proper function on the market, but the correctness of IT does not hang on the correctness of the selected effects theory *per se*.

She also explains why normal-proper functions are so important in biology: they solve a *generalization problem*. When physiologists and other biologists try to understand living systems, they're forced to make generalizations. Often enough, they make these generalizations by making a certain idealization: they envision a (possibly hypothetical) system, *all of the parts and processes of which are capable of performing their normal-proper functions*. The notion of normal-proper function is scientifically important *not* because it describes a special kind of causal power, but because it permits the generalizations that underlie biological theorizing.

The next three chapters present the core arguments for why IT is preferable to other naturalistic theories of content. It seems to me that Neander offers two main arguments. In Chapter 4, she gives her "methodological argument for IT;" it purports to show that IT is supported by considering the kinds of explanations that cognitive scientists give. But how does cognitive science support IT? Cognitive scientists postulate *normal-proper* functions. They also postulate *natural-factive* information. (This is a very thin notion of information where a state carries information about another if the former is caused by the latter. One feature of this notion of information is that there's no such thing as *mis*information.) Finally, and most importantly, they postulate brain mechanisms that have the normal-proper functions of doing things with natural-factive information. But anyone who admits this latter fact is effectively admitting the correctness of IT: "normative aboutness is posited once the normal-proper functions of cognitive systems are wedded to the natural-factive information they have the function to operate on (85)."

As a textual and interpretive matter, I think there are two different ways that we can read Neander's claim that IT is "supported by" cognitive science. Neander can be read as making either a weaker claim or, in addition, a stronger one. On this *weaker* interpretation, cognitive science supports IT in that *IT's basic ontology is not more expansive than the basic ontology of cognitive science*. Its ontology doesn't outstrip the ontology of cognitive science. That seems right to me, but that wouldn't uniquely support IT over other naturalistic theories of content. Other theories, such as consumer teleosemantics, Cummins' "picture theory," or Fodor's "asymmetrical dependence" theory, could say the same thing. (Neander does, however, argue that those alternative theories are actually *versions* of teleosemantics because they appeal, at least tacitly, to a notion of normal-proper function.)

We can also read Neander as making a stronger claim, in addition to the weaker one: that cognitive science postulates a very special *kind* of representational relation, "normative aboutness," and *it's precisely this relation that IT describes*. Here's how this argument might run. Suppose there's a brain mechanism and its function is to produce brain states that carry information about a stimulus. In other words, its function is to produce brain states that correspond well to some feature of the outside world, such as snakes. But this mechanism can fail. One way it fails is by producing that brain state when the stimulus isn't there. We now have a brain state that's (if you will) "supposed to" carry the information that (say) *there's a snake*, but *fails* to. This would be a state that depicts the world falsely. On this reading, *IT simply describes a kind of representational relation that cognitive science is already committed to*. This stronger interpretation is suggested by the following passage, among others: "in invoking the notions of natural-factive information and normal-proper function, and bringing them together to explain intentionality, [IT] only invokes what is already invoked and only brings together what is already brought together in the sciences most nearly concerned (87)."

Chapter 5 develops a second argument for IT, but to understand it properly, we need a bit of background. This chapter centers around a specific, empirically detailed example: how toads detect worms. Toads have a class of midbrain cells, the T5-2 cells, the activation of (a subset of) which is highly correlated with worms. So, let's suppose that the activation of a T5-2 cell is a representation. What, precisely, is this representation about? Naturalistic theorists of content disagree. Some argue that the right content is, *there's a worm*. Others argue that the right content is, *there's a nutritious snack*. Neander argues that the right content is, *there's a small elongated object moving parallel to its longest axis*. (Neander 1995 colorfully calls her view, with its preference for ecologically thin contents, "Low Church" teleosemantics, in contrast to "High Church" teleosemantics, which prefers ecologically thick contents: *there's a worm*; *there's a nutritious snack...*)

With this background in mind, we can now give Neander's second argument for IT. According to Neander, it can be shown on pretheoretical grounds (that is, independently of anyone's favored naturalistic theory of content), that the right content is, *there's a small elongated object moving parallel to its longest axis*. But that is precisely the content that IT gives us, and it is *not* the content that the other theories give us (as later chapters will show). Therefore, IT is preferable to those other theories.

But what exactly are these "pretheoretical grounds" that show that the right content is, *there's a small elongated object moving parallel to its longest axis*? A principle all naturalistic theorists should agree with is this: the right content ascription should hang together with the practices and norms of cognitive science. As noted above, cognitive science is committed to the idea that our ability to represent deep features of objects, like their protein content, depends on our ability to represent surface features of those objects, such as shape, size, and motion. Something in the brain, then, must represent these ecologically thin properties. The placement of T5-2 cells, moreover, suggests that those cells are sensitive solely to the latter properties. The content, *there's a small elongated object moving parallel to its longest axis*, is preferable because it points only to the kinds of properties that fit into cognitive science explanations, at least when we are talking about relatively early stages of sensory processing.

Chapter 6 rounds out her initial defense of IT by developing the notion of a *response function*. IT appeals to a very special kind of function: it says that some parts of the brain have the function of producing certain effects *in response to* certain stimuli. This deviates somewhat from the way philosophers typically think about proper functions, where functions are merely thought of as a subclass of a trait's *effects*. When we say, "the function of the heart is to circulate blood," we cite only the effects of the heart's pumping; we don't also cite its (proximal) causes, such as impulses from the vagus nerve. The purpose of this chapter is to reassure the reader that there's nothing suspect about response functions, and that, in fact, they play a prominent role in ordinary biological theorizing. For example, a function of the pineal gland is to release melatonin, but only *in response to* the dimming of light.

One of the problems that has haunted teleosemantics since its inception is the problem of "content indeterminacy." This is the subject of the final three chapters. These challenges all have the same general form: teleosemantics can't explain how we're able to make some of the fine-grained distinctions that we do, in fact, make. For example, Fodor (1990, 73) argued that if two properties, Q and C, are always co-instantiated in an organism's natural habitat, then teleosemantics can't explain how the organism represents Q, rather than C. If, in the frog's natural habitat, all flies are ambient black dots, and vice versa, then teleosemantics doesn't let us attribute to the frog's brain the content *there's a fly* rather than *there's an ambient black dot*. Here, Neander shows that there are actually six quite distinct content indeterminacy challenges, and that IT can resolve them all. Crucially, however, these last three chapters aren't just a mop-up job. In showing how IT can surmount these challenges, she's forced to develop and articulate the theory itself in more detail than she previously had. In these chapters, Neander shows that IT is actually a conjunction of three quite distinct principles: CT (Chapter 7), CDAT (Chapter 8), and the distality principle (Chapter 9).

We'll start with CT (the "causal-information version of teleosemantics"), the subject of Chapter 7. CT says that in some cases, a representation of type R has the content C because there's a neural mechanism that has the normal-proper function of causing Rs in response to Cs. A crucial qualification here is that this neural mechanism must be *causally sensitive* to the presence of Cs. For example, T5-2 activation is about things that are small, elongated, and that move parallel to their longest axis, because there's a mechanism that is causally sensitive to that configuration of visible properties and that has the function of causing T5-2 activation in response to that configuration of properties. Importantly, a system can be causally sensitive to certain properties (such as size, shape, and motion) and not others (species membership) even if those properties are always co-instantiated. Fodor's hypothetical frog actually represents ambient black dots and *not* flies, because its perceptual system is only causally sensitive to properties like size, shape, and motion, and not to species membership. With this argument, she also shows, as promised, that IT *does* deliver the pretheoretically correct content for the toad's T5-2 activity. Neander uses similar considerations to skirt the first three content-indeterminacy challenges.

Chapter 8 is devoted to two more content indeterminacy challenges. How can I represent C, rather than Q, when C is a determinate of Q, and conversely? For example, how can I represent something as being scarlet, rather than red, and vice versa? To solve them, Neander introduces the second main principle, CDAT ("causally driven analogs and teleosemantics"). CDAT says that a representation can have a content because it's produced by a mechanism that has the function of responding to environmental changes by producing that content's *analog* (195). Sometimes a perceptual system represents the world by producing a vast range of representations that differ somewhat from one another, in order to represent a vast range of stimuli that differ somewhat from one another. It does this by preserving a kind of structural similarity between the two sets. (These are sometimes called "second-order similarities.") Any one of these representations, drawn from this vast set, is called an "analog" of the corresponding member of the other set. Consider the auditory mechanisms that have the function of producing a vast range of neural representations that differ somewhat in their phenomenal loudness (as measured in decibels). These mechanisms have the function of producing representations in response to a vast range of external stimuli that differ somewhat in their sound intensity (as measured in watts per square meter).

A remarkable feature of these systems of representation is that, in principle, they would let us represent properties that *neither we, nor our ancestors, have ever directly encountered*. That is because a sensory system can produce an analog of a (possible) stimulus even if it has never actually encountered that stimulus. She illustrates this prospect by using Hume's "missing shade of blue:" Suppose you have swatches of blue that differ only in brightness, and you line them up so that you have a smooth gradation from lightest to darkest. Suppose you then remove a swatch and show the series to your friend. Presumably (Hume says), your friend would not only notice a gap, but she would be able to imagine exactly what the missing shade of blue would look like. It seems that she'd be able to do this even if, by some statistical fluke, neither she nor her ancestors ever encountered that shade of blue before. This chapter is also noteworthy in that Neander uses CDAT to begin building a theory of conceptual representations, and not just non-conceptual ones.

The ninth and final chapter deals with the problem of proximal-distal content indeterminacy, or what we can simply call the "problem of distal content." Consider our toad. We said that its T5-2 activation means, *there's a small elongated object moving parallel to its longest axis*. We took it for granted that the content was about a distal object (it's a property of the worm itself). But why not say, instead, that T5-2 activation means something like, *there's such-and-such signature pattern of retinal activation*? Both content ascriptions are licensed by her theory. IT seems to imply that T5-2 activation has *two* distinct contents, a distal one and a proximal one. But that seems wrong. We'd like to say that T5-2 activation has only to do with worms, not also with retinal impressions.

Neander's solution is this. Suppose we have a system that's causally sensitive to a distal property *by virtue of* being causally sensitive to a proximal property. Our toad's visual system is causally sensitive to the configurational property of being small, elongated, and moving parallel to its longest axis *by virtue of* being causally sensitive to special patterns of retinal stimulation. In such cases, she says, we should always prefer the most distal, rather than the more proximal, content. She also considers, and rejects, alternative solutions to the same problem, such as those that appeal to constancy mechanisms, or Dretske's (1986) appeal to a multi-modal "triangulation" principle.

Neander's book has a number of important virtues, and it should appeal to a wide range of specialists, including philosophers of biology, psychology, and mind. Here – sheerly for lack of space – I'll draw attention to two of those virtues. First, it develops a plausible and empirically-informed account of content. In particular, it represents a quite original new direction for teleosemantic theories, one that arguably overcomes – or at least indicates the path to overcoming – all of the traditional objections. Second, it can be seen as a real *agenda-setting* book. By providing the foundation for a more encompassing theory of content, one can hope that it will lure in generations of scholars who will build on it in constructive ways.

Acknowledgements

For comments on an earlier draft of the précis and response, I'm grateful to Michael Devitt, Joe Dewhurst, Fabian Hundertmark, David Papineau, Carlotta Pavese, Alex Rosenberg, Peter Schulte, and Ema Sullivan-Bissett. I also thank the critics, Frances Egan, Angela Mendelovici, David Bourget, and Chris Hill, for suggesting points of clarification. Of course, all errors and other shortcomings of the text are my own responsibility.

References

Dretske, Fred. 1986. Misrepresentation. In *Belief: Form, Content, and Function*, ed. R. Bogdan, 17-36. Oxford: Clarendon.

Fodor, J. A. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Isbell, L. A. 2009. *The Fruit, the Tree, and the Serpent: Why We See so Well*. Cambridge, MA: Harvard University Press.

Neander, K. 1995. Misrepresenting and malfunctioning. *Philosophical Studies* 79:109-141.

Neander, K. 2006. Content for cognitive science. In *Teleosemantics*, ed. George Macdonald and David Papineau, 167-194. Oxford: Clarendon Press.

Neander, K. 2013. Toward an Informational Teleosemantics. In *Millikan and her Critics*, ed. Dan Ryder, Justine Kingsbury, and Kenneth Williford, 21-40. Malden, MA: Wiley.

Pietroski, P. M. 1992. Intentionality and teleological error. *Pacific Philosophical Quarterly* 73: 267-81.

Response to Neander's Critics

Response to Frances Egan

A central claim of Neander's book is that informational teleosemantics (IT) is supported by cognitive science. In other words, IT emerges naturally when we consider how cognitive scientists explain cognition. Cognitive scientists postulate (normal-proper) functions, they postulate (natural-factive) information, and, most importantly, they postulate mechanisms that have the (normal-proper) function of doing things with (natural-factive) information. But according to IT, this is *all one needs* to explain how sensory-perceptual representations have contents. Neander infers from this that IT doesn't postulate any ingredients over and above those that cognitive scientists are already committed to, and at least in that basic sense, it gains support from cognitive science. As I noted in the précis, this is the "weaker" interpretation of Chapter 4's methodological argument for IT.

Egan raises two main challenges against Neander's view. The first is that mainstream cognitive science *doesn't* actually postulate normal-proper functions. They do postulate functions of a special kind, but the "functions" that cognitive scientists invoke aren't the same as the normal-proper functions that physiologists or evolutionary biologists invoke. If Egan were right, that would be a severe – perhaps devastating – problem for Neander's argument. Her second challenge is that, even in Neander's example of the toad's worm detection, the example that's supposed to be a paradigm case for IT, it's hard to see what explanatory work the notion of normal-proper function is doing. In fact, it's hard to see what explanatory work the very idea of representation is doing.

I'll spend most of my time on Egan's first claim, that mainstream cognitive science doesn't use the notion of normal-proper function. The first move that Egan makes, in building her case, is to claim that the specific scientific discipline at issue here is *computational cognitive science*. It's *computational cognitive science* that has no use for normal-proper functions. Computational cognitive science deploys a very different notion of "function" from physiology. When computational cognitive scientists set about understanding some cognitive achievement, such as our ability to not bump into walls, or to track flying vultures, or to size up a potential competitor for a fistfight, they explain that achievement in terms of the brain's ability to compute a *mathematical function*. An example Egan uses is edge detection. In order to extract information about edges from a barrage of retinal input, the brain appears to compute what mathematicians call a *Laplacian of a Gaussian* (LoG) function. This is a function that first reduces noise from a retinal array (the Gaussian part) and then enhances contrast (the Laplacian part). But mathematical functions, of course, aren't biological functions. The upshot is that computational cognitive science "does not depend on or presuppose the notion of normal-proper function, as [Neander's argument] requires."

There are two fairly obvious responses that someone sympathetic to Neander's view might make here. First, why does Egan think that the fate of Neander's argument hangs on the explanatory practices of *computational cognitive science* (as embodied in journals such as *Neural Computation*), rather than cognitive science more generally? "Cognitive science" is a broad, diverse field, composed of several disciplines that approach cognition from different vantage points. Some study the biological basis of cognition, and others study the abstract mathematical functions the brain computes. In fact, Neander is quite explicit that when she appeals to "mainstream cognitive science," she intends to point to a broad collection of disciplines including "neuroscience, cognitive neuroscience, and cognitive science (79)." She is even willing to grant, for the sake of argument, that some sub-branches of cognitive science might be *so* autonomous from neuroscience, *so* disconnected from the messy details of neural implementation, that they don't have much use for her normal-proper functions. (This is what she calls the "ultra-ultra autonomy stance," which she thinks is mistaken.) Still, other branches do, and that's enough to establish that the notion of normal-proper function is central to cognitive science: "To try to explain the functioning of the brain without explaining cognitive or psychological capacities would be like trying to explain the functioning of the immune system without explaining how it defends against disease (81)." So, we should not tie the fate of Neander's methodological argument to the explanatory norms of *computational* cognitive science.

Egan notes, in passing, that the reason computational cognitive science is so pivotal here is because that's the branch of cognitive science that studies cognition as *information-processing*, and therefore, that's the science that's most pertinent to Neander's argument (which, after all, has to do with information). But computational cognitive science doesn't have proprietary jurisdiction over information talk in neuroscience. Cognitive science (again, construed broadly to include neuroscience and cognitive neuroscience) is permeated with information talk, and has been since its inception. For example, Eric Kandel's famous textbook, *Principles of Neural Science*, has hundreds of references to "information" in the brain. It describes in detail the precise neural mechanisms that capture, store, transmit, transform, and utilize information about the environment. For all that, the textbook isn't about computational cognitive science.

At any rate, I don't want to spend too much time on the question of whether the soundness of Neander's methodological argument hangs on the explanatory norms of computational cognitive science. For even if we restrict our attention, as Egan urges us to do, to that subdiscipline, there is a second problem with the claim that computational cognitive science has no room for proper functions: just because computational cognitive scientists invoke *mathematical* functions, that doesn't mean they don't invoke *normal-proper* functions, too. Computational cognitive science is big enough for both kinds of function, the mathematical kind, and the normal-proper kind.

When do computational cognitive scientists invoke normal-proper functions? They do so when they're trying to figure out *which parts of the brain have the normal-proper functions of carrying out these mathematical functions*. What, for example, are the neural mechanisms that have the *normal-proper function* of computing the LoG? Marr and Hildreth, in their famous paper on vision, use neurophysiological findings to suggest that the calculation of the Gaussian part of the LoG is performed by retinal ganglion cells, and that "part of the function of one subclass of simple cells is to detect zero-crossing segments" (Marr and Hildreth 1980, 209). Here, they recognize two kinds of function, the mathematical kind and the normal-proper kind, and they think some parts of the brain have the normal-proper function of carrying out mathematical functions. This is one way that normal-proper functions are, as Neander says, "central to the functional analyses of cognition that cognitive scientists provide (73)."

Another striking example of the way computational cognitive scientists invoke both kinds of function can be found in a recent paper by Barack and Platt (2016). They begin by reminding us (as Egan has already said) that, "Neuroscientists, psychologists, philosophers, and others who study the mind invoke formal, mathematical models of behavior to characterize cognitive functions. These models are mathematical formulae that contain variables, picking out which properties of the environment the system must track, and the mathematical relations between those variables, how they must be transformed, for adaptive behavior (2)." These are Egan's mathematical functions. Then they go on to explain how *malfunction* happens: "Two distinct modes of malfunction can occur when circuit dynamics execute models of information processing. The processing models describing behavior may fail to be executed correctly by neural mechanisms, without the physiological mechanism itself malfunctioning. Or, the neural mechanisms may malfunction, thereby failing to implement the right computation (ibid)." They then speculate that psychiatric problems such as obsessive-compulsive disorder might result from the failure of a certain neural mechanism (involving parts of the ACC) to perform its function of implementing a basic exponentiation function (4). These former functions are none other than Neander's normal-proper functions. Recall that the whole point of the concept of normal-proper function is to capture two distinctions, the function-accident distinction and the function-malfunction distinction. All it is for a trait to have a normal-proper function is for some of its effects to be "accidents," (not functions), and for it to be capable of malfunction. In short, the two sorts of function, the mathematical sort and the normal-proper sort, can, and do, comfortably coexist in cognitive science.

One might suspect, in fact, that the relation between these two kinds of function is even more intimate than mere coexistence. A plausible line of thought is that the very distinction between computation and miscomputation is somehow grounded in these norms of proper functioning (see Piccinini 2015, Chp. 7; Coelho Mollo 2019, for discussion). Miscomputation is somehow *parasitic* on failure of function; the former cannot exist without the latter. According to this line of reasoning, what it is for a neural mechanism to *miscompute* a certain function, such as the Gaussian function, is, in part, for the mechanism to *fail to perform* its normal-proper function of computing the Gaussian. If that line of reasoning is correct, then there would be a kind of absurdity in the proposition that cognitive scientists could postulate computational functions without also, at least implicitly, postulating normal-proper functions, too.

Egan seems to anticipate this line of thought, and she forcefully attempts to rebut it. As she tells us, once we have the idea of a mapping, we can make a perfectly cogent distinction between a device *computing correctly* and *miscomputing*. That distinction does not essentially require appeal to normal-proper functions. Egan's idea seems to be this. The notion of "normal functioning" at play in computational cognitive science arises in a "distinctive way," a way that is different from how it arises in, say, physiology. In the former, for a physical system to compute a mathematical function F, there must be a *mapping* from physical states of the system to the values of F, such that the causal relations between the states of the system mirror the mathematical and logical relations between the values of F. For example, she asks us to consider an *adder*, a device that computes the addition function. When the adder goes into a physical state specified by the mapping as *n*, and then goes into the physical state specified as *m*, it goes into the state specified as *n*+*m*.

Armed with the concept of a mapping, we can say what it is for a physical system to "compute correctly" or to "miscompute" with respect to F (say, the addition function): our adder computes addition *correctly* when it adds two numbers, and it *miscomputes* when it does not add two numbers. Hence, she claims, "In attributing a mathematical capacity or *competence* to a physical system – to compute vector displacement, or a smoothing function – a computational model supports attributions of *correctness* and *mistake*." We can make such attributions without postulating a deeper notion of normal function.

I think Egan's argument, however – that the bare notion of a mapping gives us all we need to make cogent attributions of correctness and mistake – doesn't work. This can be illustrated by a simple example. Suppose someone gives me a pocket *multiplier*. Its function is to take two numbers as input and give the product of those two numbers as output. (To help flesh out the example, suppose that I don't *know* it's a pocket multiplier, and I mistakenly think that it's a pocket *adder*, because, say, the multiplication symbol has worn off.) Now, suppose that each time I punch in the symbol "7," followed by the illegible symbol, followed by "3," followed by "=," it outputs the symbol "21." Is this device miscomputing the addition function?

It seems to me that the answer is *no*. The device is not miscomputing the addition function, and that's because the device does not have the (normal-proper) *function* of computing the addition function. It has the function of computing the multiplication function. If it did, however, have the (normal-proper) function of computing the addition function, then it would be miscomputing the addition function – but not otherwise. It is in this sense that the distinction between computing correctly and miscomputing appears to depend on a deeper, independent notion of normal-proper function.

As noted above, Egan has a second criticism of Neander: even for the example that's supposed to be most favorable to Neander's view, concerning the toad's ability to detect worms, it's hard to see what theoretical labor normal-proper functions are actually performing. Consider the midbrain cells involved in prey-detection. In explaining the toad's prey-recognition capacity, "the theorist must (1) isolate the neural structures that play the appropriate role in mediating the prey-recognition process (T5-2 cells), and (2) specify precisely the conditions of their activation (a moving worm-like stimulus)." But once you've done both of those things, you've done all the hard theoretical work! To say those cells have the "normal-proper function of producing that response" would seem to be a useless appendage to the theoretical labor that's already been done.

This critique raises a much broader concern that people have raised with respect to normal-proper functions and, in particular, with selected effect functions. I'll discuss the problem as it arises for selected effect functions and then show how it generalizes to normal-proper functions. The criticism is based on the fact that selected effect functions are, in some sense of the term, "epiphenomenal" (as Mossio et al. 2009, 821 put it). When a system comes to acquire a new selected effect function, it doesn't necessarily come to acquire any new causal powers. The first time a zebra used its stripes to deter biting flies, its stripes didn't have the *function* of doing so. They only came to acquire the *function* of deterring biting flies after several rounds of selection, but – after selection – they didn't necessarily deter flies any better than they did before. So it's tempting to think that selected effect functions don't play any explanatory role in science.

As Neander points out, the problem generalizes. It's not just selected effects functions that are epiphenomenal, but *any coherent account of normal-proper function will make functions "epiphenomenal."* Normal-proper functions and causal powers *must* diverge (58). That's because a trait with a normal-proper function can malfunction. Something malfunctions when it has the function to F but it lacks the causal power to F. For example, even if the pineal gland has the normal-proper function of producing melatonin, that doesn't mean any particular pineal gland (mine or yours) has the causal power to produce melatonin.

What, then, is the explanatory role of normal-proper functions? Why aren't they just a useless appendage that biologists appeal to after they've already done the hard, theoretical work? The short answer, according to Neander, is that steps 1 and 2 that Egan outlines already presuppose a notion of normal-proper function. The concept of normal-proper function is a necessary starting point for carrying out the tough theoretical work we're talking about, not an afterthought, and that in two ways. *First*, the notion of normal-proper function plays a role in helping physiologists solve what Neander calls a "generalization problem." The generalization problem is the problem of making reliable generalizations about complex living systems despite the extraordinary variability they exhibit. Physiologists solve this problem, she holds, by describing idealized systems in which each part performs its function exactly as it's supposed to. That does *not* imply that biology imposes some monolithic or oppressive ideal of "species normalcy." As Neander emphasizes, there are different ways to be "normal," even within a single species (63).

Here's a *second* way that neuroscience presupposes normal-proper functions. When we try to assess a cell's response profile, for example, when we say, "this class of retinal ganglion cell (ON-center/OFF-surround) responds best when only the center of the receptive field is illuminated," what we're really saying is, "when these cells are *functioning normally*, they respond best when only the center of the receptive field is illuminated." Presumably, those cells would respond even *better* by impaling their soma with a pipette and jolting them with short bursts of electricity. But we don't consider that situation when we're trying to figure out their response profiles, since that doesn't represent its *normal functioning*. (See Neander's comments on Fodor, p. 92, for similar considerations.) Making good generalizations about cells and their response profiles requires a prior notion of proper function; it's not an afterthought.

Egan points out that the same sort of redundancy argument can be leveled not only against normal-proper functions, but against representations, too. What's the point of calling a brain state a "representation?" To be clear, she *does* see the theoretical virtue of attributing representations at the more sophisticated, *conceptual* level. When I say, "Makayla thinks that the Koenigsegg is more beautiful than the Maserati," that helps me to make good generalizations and predictions about Makayla, even if I don't know anything about the messy and idiosyncratic neurobiological facts that implement her belief. But it's hard to see why calling T5-2 activation a "representation" is scientifically useful at all. This is a concern that's been raised even by those who are friendly to teleosemantics.

There are several responses one might make to this concern. First, much of what Neander says about normal-proper function can be said about representation, too. Even calling the toad's visual system a "prey-recognition process," as Egan does, is an intentional description. It describes the system as one that's supposed to *recognize prey*. In this manner, using intentional descriptions helps neuroscientists solve a generalization problem. It helps them classify the parts and processes of the toad's brain into meaningful, species-typical, biological units. Second, Neander spends quite a bit of time, in Chapter 2, documenting how vision scientists postulate error-permitting, non-conceptual representations, and how it would be almost impossible to describe some of the pathologies that we do, in fact, encounter, without postulating the existence of nonconceptual representations that misrepresent how the world is.

But there's a third, and deeper, response. If representation *can* be given a correct naturalistic, and reductionist, definition ("X is a representation of Y just in case X bears naturalistic relation R to Y"), then we should expect that there are some fairly simple neural systems for which using the concept of a representation isn't doing any more work than merely using the concepts that make up its *definiens*, that is, the concepts designated in the right-hand side of our definition. That's not a flaw; it's a feature. That is how good reductive definitions work. As Neander says, "…we paint the success of a naturalistic theory of mental content in the colors of failure if we demand of the representational posit, even at its most elementary, that it have more explanatory heft than what it supervenes on (that in terms of which it is naturalized)." (88)

Response to Angela Mendelovici and David Bourget

Mendelovici and Bourget develop an argument against IT that appeals centrally to the idea that we have direct, introspective access to the contents of our own representations. Roughly, they argue that IT, along with some fairly plausible empirical facts, entails that color representations represent surface reflectance properties of objects, or something along those lines. But, they argue, it's introspectively *obvious* that color representations aren't about surface reflectance properties of objects. Therefore, IT is false (see Mendelovici 2018, 41, for further discussion).

Of course, this argument would appear to be straightforwardly question-begging – not in the formal sense ("P, therefore P") – but in the broader sense that it asserts, or assumes, something that their opponents would never accept. Teleosemanticists, and I suppose many people who accept naturalistic theories of content, don't think we have direct, introspective access to the contents of our representations. That's because, as I'll elaborate below, the contents of our representations are determined by history, and not by our current psychological or phenomenological states. We can't just intuit or *see* them. That doesn't mean introspection has no place in a mature teleosemantics, but that we don't take its claims at face value. So, Mendelovici and Bourget's argument won't, and shouldn't, convince anyone who's favorably inclined toward teleosemantics. Still, their paper engages in an interesting way with some quite foundational issues in thinking about intentionality. To get our bearings, it's worth taking a moment to reconsider some of the main virtues of IT, so that we can put their argument from introspection in its proper light.

Teleosemantics has at least five general virtues. The first virtue is that it's naturalistic, and that in *two* senses: it does not invoke any entities or processes that transcend the natural world, and it presumes that intentionality can be entirely understood, explicated, or "reduced to," non-intentional states and processes. Nobody has expressed this aim of naturalistic theories of content better than Fodor: "If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else (1987, 87)." A second virtue is that, as Neander emphasizes, it gains support from the explanatory norms and practices of contemporary cognitive science.

A third consideration in its favor is that it's taxonomically general. It applies not only to human representational powers, but to the representational powers of other animals, such as vervet monkeys, voles, and toads. Some of these creatures, like toads, possibly have no phenomenal consciousness to speak of. They certainly have no introspective access to the contents of their own representations. It also applies to the representational powers of sub-personal cognitive states. At least *at the outset*, a good theory of content shouldn't tie intentionality too closely to phenomenal consciousness, introspective access, or the like. This consideration is particularly important for those of us who hold out hope of giving an *evolutionary* account of intentionality. If Darwin was even remotely correct in his basic vision of life on earth, then it stands to reason that human intentionality somehow builds on the (quasi-) representational powers of simpler creatures. As Neander notes: "Creatures slid, scuttled, and swam in the post-primordial sea, perceiving their surroundings and hunting and hiding and so on for millions of years before our ancestors managed to crawl onto land, let alone stand upright and eventually start conversing (13)."

A fourth consideration in favor of teleosemantics is that it yields a straightforward account of misrepresentation. Misrepresentation involves the failure of a system to perform its natural function. Other theories of intentionality have struggled to make sense of how a representation can ever be incorrect. Millikan (2004, 63) actually goes so far as to say that teleosemantics is *just* a theory of misrepresentation – though perhaps that's going a bit *too* far. It also accounts for other features of intentionality, such as its "directed" character. The directed character of intentionality comes from the directed character of functions.

Finally, teleosemantics ties intentionality very closely to the functional role of representation. It helps us to understand one way that attributing contents to some mental states or brain states can be scientifically illuminating. When we understand that T5-2 activation in the toad's midbrain means something like, *there's a worm in front of me*, we can grasp the sheer *appropriateness* of the behaviors that T5-2 activation induces, namely, tongue-snapping. If you're a toad and you have a T5-2 activation, snapping your tongue is usually a great thing to do, since *T5-2 activation represents worms*. (There are some complications here for Neander's view because she thinks the right content for the toad is, *there's a small elongated object moving parallel to its longest axis*, and it isn't immediately obvious, without any background ecological knowledge, why snapping is the right thing to do in that situation. But someone with a modicum of background ecological knowledge *would* be able to see, from that content ascription alone, why snapping *is* the right thing to do.) Teleosemantics can account for the psychological role of some representations, too. For example, when I have a representation of a snake, that representation causes me to feel panic and think about fleeing. One consequence of teleosemantics is that if you're representing something as a snake, rather than a dog or canoe, it's very likely that you will exhibit snake-appropriate thoughts, feelings, and actions.

Of course, another consequence of teleosemantics is that we don't, as a rule, have some sort of direct, unmediated, introspective access to the content of our own representations. That is because teleosemantics, like many naturalistic theories of content, is externalist. Whether the thought I'm currently harboring is a thought about a coastal taipan, or a thought about an elegant lobulia, depends on deep facts about my evolutionary history, and to some extent on the course of my individual development. It's not something I can just *see*. Ultimately, whether one is willing to accept teleosemantics depends on whether one thinks these theoretical virtues outweigh the untutored claims of introspection.

This is a point, of course, that early teleosemantic theorists like Millikan and Dennett emphasized, with unmatched rhetorical flourish. Millikan denounced the (in her mind) false and pernicious doctrine that she called "meaning rationalism," arguing instead that "we are no more in a position to know *merely via Cartesian reflection* that we are truly *thinking*, i.e., that we or our thoughts intend anything, than that we are thinking truly (1984, 92)." Dennett (1987, 313) put the point even more starkly, if such a thing were possible: "Either you must abandon meaning rationalism – the idea that you are unlike the fledging cuckoo not only in having access, but also in having privileged access to your meanings – or you must abandon the naturalism that insists that you are, after all, just a product of natural selection…"

This doesn't mean that teleosemantics *forbids* introspection from yielding some information about the intentional properties of our own mental states, particularly when we get to fairly complex conceptual representations. Neander (2012) was open to the possibility that we enjoy "introspective access to conceptual structures," but she also seemed to think that such introspective access might be a feature of conceptual, rather than nonconceptual, representation. There's nothing in teleosemantics that makes it *impossible* that human beings might have evolved in such a way that we sometimes have correct (second-order) thoughts about the contents of our (first-order) thoughts. This would be particularly valuable in social settings, where people constantly ask each other, "what are you thinking about?" (Of course, teleosemantics is also consistent with the possibility that we've evolved the capacity to systematically *misrepresent* the contents of our first-order thoughts, for example, in settings where an ounce of self-deception could further our aims and goals.)

That's enough by way of an overview of teleosemantics' virtues. Now on to the details of Mendelovici and Bourget's arguments. Their main conclusion is that IT fails as a theory of intentionality, because it cannot account for the content of color representations. Their argument is fairly complex, so I'll give a schematic overview and then expand it. There are three premises.

> 1. According to Neander's view, color representations have their contents in one of two ways, by CDAT, or by CT (I'll remind the reader shortly of what those abbreviations mean).

2. Color representations don't get their contents by CDAT, because color representations aren't analogs, in Neander's sense, of anything in nature.

3. Nor do they get their contents by CT, since that would have the obviously false implication that color representations are about (e.g.) surface reflectance properties.

Therefore,

4. Neander's theory can't explain the content of color representations.

I think premise 3 is question-begging, but we'll go through each in turn. Premise 1 simply notes that, for Neander, there are two quite different ways that a representational vehicle (a specific pattern of neural activation, say) can acquire its content. First, it can acquire a content by virtue of CT, that is, by virtue of there being a mechanism that has the function of producing *that* sort of representation in response to *that* sort of content. (There is a mechanism that has the function of producing a certain pattern of neural activation in response to snakes, in which case that pattern of activation has the content, *there's a shiny, slithering, S-shaped stimulus*, or something like that.) Second, it can acquire a content by virtue of CDAT, that is, by virtue of there being a mechanism that has the function of producing an *analog* of a certain sort of property in the world. In this case, there's a mechanism that has the function of producing a diverse array of representations, the similarities and differences between which mirror the similarities and differences in some feature of the world. A virtue of CDAT, as I noted in the précis, is that it would let us represent properties that neither we, nor our ancestors, have ever actually encountered (recall Hume's missing shade of blue). The auditory mechanism that's responsible for encoding information about sound intensity (measured in watts per square meter), and which is associated with the experienced loudness of sounds (measured in decibels), is probably such a mechanism. Even if, implausibly enough, I've never encountered a sound intensity of 10^{-8} watts/m^2 (a loud dishwasher), I can form a representation of it, so long as I've experienced intensities in its neighborhood.

Premise 2 points out something that Neander explicitly suggests: color representations (that is, the neural patterns of activation associated with phenomenal colors) probably don't get their contents in this second, analog way, by CDAT. That's because color representations don't seem to be analogs of anything in the outside world Mendelovici and Bourget call this the "structural mismatch" problem, to be contrasted with the "qualitative mismatch" problem, below. The patterns of similarity and difference that obtain amongst our color representations don't mirror the patterns of similarity and difference among, say, the surface spectral reflectance (SSR) of different objects. For example, consider phenomenal red, violet, and yellow. Phenomenal red is more similar to phenomenal violet than it is to phenomenal yellow. Presumably, that's because the neural underpinnings of phenomenal red are more similar to the neural underpinnings of phenomenal violet than they are to the neural underpinnings of phenomenal yellow. Surprisingly, however, the SSR associated with phenomenal red is more similar to the SSR associated with phenomenal *yellow* than it is to the SSR associated with phenomenal violet. This and other examples show that there is a "structural mismatch" between color representations and SSRs which prevents color representations from being analogs of SSRs. And the same point holds for other naturalistic properties, other than SSRs, that we might pick out.

We should pause to note that Neander actually takes the existence of this "structural mismatch" as a working hypothesis, namely, "that there are no sufficiently ordered and systematic analogous relations to be found among the surface reflectance properties of external objects that cause color sensations in us (198)." In other words, she appears to be willing to accept premise 2; nor do Mendelovici and Bourget say otherwise. We should also note that this claim, that color representations *aren't* analogs of anything in nature, is actually quite controversial. Paul Churchland (2007), whom Neander cites, develops an ingenious argument that the phenomenological color space maps quite neatly onto a space defined over entities that are mathematical functions of surface reflectance profiles.

If color representations don't get their contents by being analogs of something in the world, they must get their contents in this simpler way, from CT. Let's suppose then, for the sake of argument, that color representations are about SSRs, because there are some evolved mechanisms that have the normal-proper function of producing those representations in response to SSRs, despite the fact that those representations aren't analogs of anything in nature. All of this sounds reasonable so far.

So, what *exactly* is the problem? For Mendelovici and Bourget, the problem is this: *color representations obviously aren't about SSRs* (this is their "qualitative mismatch" problem). Here is what they say: "As noted above, color representations are not analogs of kolors [i.e., SSRs]. *It is even more obvious that color contents are not analogs of kolors.* For example, the content of a representation of a particular shade of red is more similar to the content of a representation of a particular shade of violet than it is to the content of the representation of a certain shade of yellow, but the kolor properties assigned by Neander's theory to the representations of red and violet are less similar to one another than either is to the kolor property assigned to the representation of yellow (my emphasis)." Later, they're even more explicit about this commitment: "Introspection reveals contents involving bluish, reddish, and other color qualities that are in no way captured by content attributions in terms of kolors – i.e., surface reflectance profiles – or related physical features of putatively colored objects." This is the argument that's question-begging, since it assumes that we have a direct, introspective, grasp of the contents of our color representations, a grasp that teleosemanticists uniformly deny.

In short, it's not obvious to me that color representations aren't about SSRs. There are, however, some truths in this neighborhood that strike me as rather obvious, and I suspect that Mendelovici and Bourget might have latched onto one of those, and mistakenly construed it as an obvious truth about color contents. Here is one obvious truth: color representations (which, again, we are taking to be brain states) are associated with phenomenal qualities. There's *something it's like* to relish the deep lavender of Rothko's *White Center (Yellow, Pink and Lavender on Rose)*. Neander never denies that. She remains fairly agnostic on the question of how certain representational vehicles come to be associated with certain phenomenal qualities, and speculates that those qualities are somehow determined by the "shape" of the representational vehicles (4). She just doesn't think that the phenomenal quality associated with a representational vehicle is the same thing as its content.

Here is another obvious truth: human beings can form *concepts* about the phenomenal experiences they enjoy. They can say (and think) things like, "I adore the way baby blue looks in this room." This, quite possibly, involves a concept about a phenomenal color. But Neander, at least in this book, doesn't take a position one way or the other on the contents of *concepts*. Here, she's interested primarily in the contents of nonconceptual representations.

One might wonder if Mendelovici and Bourget give any *non*-question-begging arguments against IT, that is, arguments that might give someone serious pause if they're favorable to IT, or teleosemantics more generally. From what I can tell, there are two such arguments. The first is that IT doesn't match the *psychological role* of our color representations. (One might see this point as another way of arguing for premise 3, rather than an entirely separate argument.) Color representations have a number of systematic psychological effects: "a perceptual experience of a red tomato before you is likely to lead to a belief that there is a red tomato before you, the higher-order thought that you are perceiving a red tomato, and tomato-appropriate behaviors…" They argue that our color representations do not have the sorts of systemic psychological effects they should have if they were about SSRs: "Likewise, our color representations do not behave as if they represented kolors [i.e., SSRs]: they do not cause beliefs about particular kolors, higher-order thoughts about representing kolors, or behavior aimed at such kolors."

I agree that color representations don't typically give rise to concepts about SSRs, since most people don't have the concept of an SSR. But they do cause thoughts and actions that are appropriate to SSRs. This is all but guaranteed by the way that teleosemantics is set up. According to any version of teleosemantics, the reason we have representations of SSRs at all is because *there's a right (and a wrong) way to think and act when confronted by different kinds of surfaces*. Being able to detect, at a glance, the presence of different kinds of surfaces can be vital to staying alive. Suppose, for example, you're looking for red raspberries that are partly obscured by green foliage. The ability to represent different SSRs is what enables us to readily discern that there are two quite different types of surfaces in front of us – and thereby hopefully grasp a handful of berries, rather than thorns.

The second argument is that, if color representations aren't analogs of anything in the natural world, then IT can't explain how someone could ever have a representation of a color that neither they, nor their ancestors, have ever personally encountered (Hume's "missing shade of blue"). The problem raised here is sometimes known as the problem of novel contents, and many philosophers have worked on it for teleosemantics. Even if Neander's preferred solution in terms of analog representations doesn't actually work for color (as she suspected – see p. 201), there are other solutions that might work. Millikan's preferred solution is in terms of direct and derived proper functions (see Kingsbury 2006 for a lucid overview of how this distinction can be used to understand novel contents). Garson and Papineau (2019) recently develop the idea that we can explain novel contents by appealing to the workings of ontogenetic selection processes. Hundertmark (2019) explains novel contents in terms of the idea that natural selection selects for "complex," rather than "simple," dispositions. If anything, there are too many solutions to the problem of novel contents, rather than too few.

Mendelovici and Bourget raise a final critique of Neander's worm-detection example. Recall that Neander wanted to provide a theory-neutral or "pretheoretical" assessment of what the correct content ascription for the toad's T5-2 activation is. Is it *there's a worm*? Or *there's a nutritious meal*? Or *there's a small elongated object moving parallel to its longest axis*? Neander argues on "pretheoretical" grounds that the correct content ascription is the last, because that is the one that hangs together with the explanatory role of content in cognitive science. Cognitive scientists attempt to explain our ability to represent "deep" properties of objects in terms of our ability to represent "shallow" properties, such as size, shape, and motion.

Mendelovici and Bourget state that this is not a theory-neutral assessment of content, but rather, that it seems to "outright assume" the correctness of some causal-informational notion of intentionality. That is not entirely correct. By chapter 5, when she approaches the problem of toad vision, she does assume that the right theory of content is a naturalistic theory informed by cognitive science and evolution. This was a point that the earlier chapters tried to motivate. By the time she reaches the midpoint of her book, she is no longer trying to prove that. Rather, as I understand her, she is trying to adjudicate what Egan calls an "in-house debate" amongst teleosemanticists (and others who offer up naturalistic theories in the same spirit, such as causal or "asymmetrical-dependence" theories of content). It's not intended to adjudicate between the teleosemanticist and the proponent of, say, a phenomenological or even a theistic account of intentionality. She tells us at the outset of the chapter that her goal is to figure out the correct content, *conditional on* the assumption that "a mainstream information-processing approach to explaining cognition [is] on the right lines in key respects" (97).

Response to Christopher Hill

Hill purports to discover a serious inconsistency in Neander's book, and offers some friendly recommendations about how she might resolve that inconsistency. But the alleged inconsistency, I think, doesn't exist. So, while Hill's comments are quite interesting and worthy of reflection in their own right, particularly his innovative approach to thinking about distal content, the theory Neander outlines isn't in need of help in this particular regard.

The alleged inconsistency centers around Neander's appeal to what Hill calls a "discriminability thesis," which, he says, is the "third pillar" of Neander's theory, along with CT and CDAT. So, he sees Neander's theory as a three-pronged view that consists of CT, CDAT, and the discriminability thesis (or perhaps he sees it as a four-pronged view that consists of CT, CDAT, the discriminability thesis, and the distality principle). The discriminability thesis is this: if an organism's sensory system can't discriminate between A and B, then it doesn't have (sensory-perceptual) representations with the content, *there's an A*. If I can't tell the difference between a pinot noir and a gamay noir, then I don't have a (sensory-perceptual) representation with the content, *there's a pinot noir* (nor do I have a representation with the content, *there's a gamay noir*). If anything, I have a representation with the content, *there's that acidic red stuff with strong berry notes*, or something like that. This discriminability thesis seems to motivate her conclusion that T5-2 activation means, *there's a small elongated object moving parallel to its longest axis*, rather than, *there's a nutritious snack.*

The problem Hill notes is that, if we apply the discriminability thesis consistently, it seems to yield quite unattractive consequences for Neander's project. Let A and B represent distal and proximal links in a cause-and-effect chain (say, a worm and a retinal impression of a worm). Let’s suppose that an organism can’t discriminate between them. Then it wouldn’t have a representation that means, *there’s a worm*, rather than, *there’s a worm-shaped retinal impression* – contrary to Neander’s claim that sensory-perceptual representations are about distal, not proximal, properties. (As Hill rhetorically asks, “can we really discriminate between retinal images and their distal causes?”) Hill concludes that, to avoid this embarrassing implication, Neander applies her discriminability thesis *inconsistently*, according to something like the following rule: when the thesis yields the results I like, then apply it; when it yields the results I don’t like, withhold it.

The problem with this assessment is that Neander is not committed to anything like an unrestricted discriminability thesis. I find little evidence that this discriminability thesis represents a third pillar of her thought, alongside CT and CDAT. From what I can tell, she raises the issue only once, in what appears to be a somewhat incidental remark about twin-earth cases (120). (This is the passage Hill cites.) It is true that a *restricted* discriminability thesis is implied by CT and CDAT, but the restricted version thus implied doesn’t bear on the problem of distal content, as I will explain.

This, instead, appears to be the overarching structure of her theory. In order to figure out what a (sensory-perceptual) representation represents, we *first* apply CT (the non-analog version) or CDAT (the analog version), whichever is appropriate in the context. Then, if CT/CDAT yields *too many contents*, we apply her distality principle to whittle them down to a manageable size. That’s the logical (and admirably consistent) structure of her theory. It is *not*: in order to find out what a representation represents, apply CT/CDAT, *along with* the discriminability thesis. Then, when the discriminability thesis yields intuitively bad results, forget about the discriminability thesis and apply the distality principle instead.

So how, exactly, does a restricted discriminability thesis emerge from Neander’s theory? Let’s focus on CT, rather than CDAT, since the point can be made just by looking at CT. Here is Neander’s definition of CT: “A sensory-perceptual representation, R, which is an (R-type) event in a sensory-perceptual system (S), has the content *there’s C* if and only if S has the function to produce R-type events in response to C-type events (in virtue of their C-ness) (151).” A crucial implication here (based on the way that she analyzes the notion of a response function) is that *a sensory system can only represent properties it’s causally sensitive to*. What exactly does that mean? It means that I can only represent those properties of an object that figure into a cause-and-effect chain that culminates with the stimulation of my sensory receptors. Which properties are those? According to cognitive neuroscience, the properties of an object that stimulate my sensory receptors include motion, shape, size, texture, thermal properties, and so on. They do not include properties such as genus membership (e.g., the property of being a member of the genus *Lumbricus*) or protein content. So, *CT alone*, without help from an added discriminability thesis, is what leads Neander to conclude that sensory-perceptual representations are ecologically “thin;” they are about surface features of objects.

Here’s one way to describe this implication of CT: if the presence of one property rather than another doesn’t have any impact on a creature’s sensorium, then that creature doesn’t have a (sensory-perceptual) representation of that property. If a toad can’t visually tell the difference between a common earthworm and a European nightcrawler, then it can't represent something as being a common earthworm. Hence, CT *does* imply a discriminability thesis, at least in the ordinary case where we're talking about objects that are equally distal. And that's what Neander appears to get at in her aside to Millikan about water and twin-water:

> To be clear, what I deny is not the possibility that a nonconceptual representation might have the content water/H2O as distinct from twin water/XYZ. Rather, I deny that there could be a nonconceptual representation that has such content despite no ability, on the part of any member of the species, to distinguish between the two. In other words, an attenuated form of verificationism is quite right for nonconceptual as opposed to conceptual representations. (120)

Contra Millikan, if I can't tell the difference between H2O and XYZ, then I don't have a (sensory-perceptual) representation with the content, *there's some H2O*.

Note, however, that the restricted discriminability thesis that's implied by CT doesn't say anything about our ability to represent distal *versus* proximal stimuli. For consider two stimuli: a small, elongated worm, and a corresponding pattern of retinal activity. According to CT, which of these two properties does the toad represent: the worm's property of being small and elongated, or the retina's property of having such-and-such pattern of activation? It represents whichever one it's causally sensitive to. But which one is that? The answer is *both*: it's causally sensitive to the worm's being small and elongated, and it's also causally sensitive to the retina's having such-and-such pattern of activation. In fact, *it's causally sensitive to the worm's being small and elongated, by virtue of being causally sensitive to the retina's having such-and-such pattern of activation* (220). CT/CDAT alone doesn't privilege one content over the other. That's why we need additional principles.

This is where the distality principle kicks in (222). Suppose CT picks out several different properties that can be arranged in a distal-to-proximal sequence. The distality principle says that the most *distal* of these properties is the one the system represents. (Her rendition of the principle is a bit more complicated, but it's the same idea.) Of course, critics are free to ask why the distality principle is the *right* principle. One might think it's incorrect, or that it yields further problems for Neander – see Price (2014), Artiga (2015), Schulte (2018), and Garson (2019, Chapter 12) for discussion. Peter Schulte, for example, is among those philosophers who accept much of IT, but thinks the distality principle is mistaken. He develops a very clever thought experiment to try to debunk it. Suppose there's a bug, and whenever the bug is filled with potassium, it turns red. Suppose there's a frog whose visual system is designed to detect the redness of bugs, and it's designed that way because frogs need the potassium that red bugs contain. In this example, our fictitious frog is causally sensitive to the potassium-richness of bugs *by virtue of* being causally sensitive to the redness of bugs. So, Neander's distality principle, he claims, actually has the consequence that the frog's sensory system represents the content, *there's that potassium-rich thing*. But that would contradict her claim that such representations are only about surface features of objects. Schulte favors the idea that we appeal to constancy mechanisms, rather than the distality principle, to solve the problem of distal contents, though both Hill and Neander are skeptical about that approach. From what I can tell, these issues are quite ripe for further philosophical exploration.

References

Artiga, Marc. 2015. Review of *Millikan and her Critics*. *Mind* 124: 679-683.

Barack, D. L., and Platt, M. L. 2016. Neurocomputational nosology: Malfunctions of models and mechanisms. *Frontiers in Psychology* 7: 602.

Coelho Mollo, D. 2019. Are there teleological functions to compute? *Philosophy of Science* 86: 431-452.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Fodor, J. A. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Garson, J. 2019. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

Garson, J., and Papineau, D. 2019. Teleosemantics, selection and novel contents. *Biology and Philosophy* 34:36.

Hundertmark, F. 2019. Explaining how to perceive the new: Causal-informational teleosemantics and productive response functions. *Synthese*. https://doi.org/10.1007/s11229-019-02406-3.

Kingsbury, J. 2006. A proper understanding of Millikan. *Acta Analytica* 21: 23-40.

Marr, D. and Hildreth. E. 1980. Theory of edge detection. *Proceedings of the Royal Society B* 207: 187-217.

Mendelovici, A. 2018. *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. 2004. *Varieties of Meaning*. Cambridge, MA: MIT Press.

Mossio, M., Saborido, C., and Moreno, A. 2009. An organizational account for biological functions. *British Journal for the Philosophy of Science* 60: 813–841.

Neander, K. 2012. Teleosemantic theories of mental content. Stanford Encyclopedia of philosophy. http://plato.stanford.edu/entries/content-teleological/

Piccinini, G. 2015. *Physical Computation: A Mechanistic Approach*. Oxford: Oxford University Press.

Price, Carolyn. 2014. Teleosemantics re-examined: Content, explanation, and norms. *Biology and Philosophy* 29: 587-596.

Schulte, P. 2018. Perceiving the world outside: How to solve the distality problem for informational teleosemantics. *The Philosophical Quarterly* 68: 349-369.