Representation in Cognitive Science, by Nicholas Shea. Oxford: Oxford University Press, 2018. Pp. 292.

A central component of the cognitive revolution is commitment to explaining behavior by reference to internal representations of the world. This core aspect of cognitive science raises a number of theoretical questions. For one thing, how, exactly, does invoking representations help us make sense of behavior? Further, given that a representation by its very nature is able to serve as a stand-in for something else not immediately present, how do we determine what a representation stands for or means? In this exceptional contribution to the philosophical literature on representation, Nicholas Shea addresses these fundamental issues within a teleosemantic framework. His book lies in the tradition of Ruth Millikan's *Language, Thought and Other Biological Categories* and Fred Dretske's *Explaining Behavior: Reasons in a World of Causes*. Not only is Shea developing a naturalistic version of teleosemantics in the spirit of these classic texts; like Millikan and Dretske, he has provided a shining example of what philosophy at its best can achieve.

Among the most significant accomplishments of the book are Shea's detailed responses to the two questions mentioned above. First, the book offers a lucid defense of the explanatory power of representations in cognitive science. It does so by clarifying how appealing to representations allows us to offer better explanations of behavior than would otherwise be available and by addressing some of the more pressing skeptical worries about the idea that representations help explain behavior, including doubts raised by William Ramsey in his influential book *Representation Reconsidered*. Second, by offering the most compelling version of teleosemantics to date, Shea has considerably advanced discussion of how best to characterize the meaning or content of representations.

One of the ways Shea manages to advance discussion of long-standing debates is by circumscribing the form of representation at issue. He is concerned with representations as they figure in scientific explanations of behavior. Generally speaking, cognitive scientists invoke representations in order to explain why animals succeed or fail in various goal-directed activities. Often scientists explain these successes and failures by appeal to neural representations at a *subpersonal* (i.e. non-conscious, non-doxastic) level. By restricting

attention to subpersonal states invoked in scientific theories, Shea is able to set aside whatever intuitions we may have about representational content and focus on actual cases where cognitive science has managed to shed light on behavior by invoking neural representations. With case studies as his guide, Shea makes headway on debates which had previously stalled due to conflicting intuitions.

But the most significant way Shea has altered the focus of debate on representational content is by rethinking what, exactly, representational content is supposed to explain. The broad idea is the familiar one just mentioned: appeal to representations allows us to make sense of successes and failures of goal-directed behavior. Shea calls these behaviors or outputs of a system *task functions* of the system. Task functions can take a variety of forms. They might be trained responses in the psychophysics lab, behavioral traits observed in the wild by behavioral ecologists, an artefact's intended outputs, and so on. Shea's focus is on outputs of organisms, and the ways the outputs of organisms can count as goal-directed, i.e. as functions of organisms, even in the absence of goals set by person-level beliefs and desires. Of course, not all goal-directed behavior of organisms is to be explained by reference to representations. Much of what is distinctive about Shea's approach to representation lies in his insightful remarks on exactly which forms of goal-directedness are to be explained in representational terms.

According to Shea, behavior explicable in representational terms needs to be robustly goal-directed. What makes a system's output a robust function of the system is the fact that the system produces the output in response to a variety of inputs and across an array of external conditions. For example, a predator hard-wired with an avoidance response to the warning coloration (aposematic signal) of poisonous dart frogs might successfully avoid the frogs even though the colors vary across individual frogs and even though viewing conditions change throughout the day. Shea makes use of this robustness requirement in addressing a familiar worry due to William Ramsey, who notes that the job description of a representation needs to go beyond merely serving as a link in the causal chain extending from stimulus to behavior. Otherwise we end up with an overly permissive or liberal view about which things count as representations. Shea's proposal is that representational systems differ from causal mediators insofar as they are decoupled from specific inputs and outputs (cf. Sterelny 2003, Burge 2010, and Schulte 2015). This decoupling is secured by the robustness requirement.

A second requirement must be satisfied in order for system output to be amenable to representational explanation: the output needs to be goal-directed as a result of a stabilizing process. Recall that representations are supposed to help explain successes and failures in a system's goal-directed behavior. Accordingly, there needs to be some fact about what a system is supposed to do, the system's goal or task function. Shea defends the orthodox view that a behavioral trait has a function in virtue of its history. The most familiar application of this view has us look to natural selection as the stabilizing force which fixes the function of the output. The function of an organism's output is determined by the contribution that type of output made to the inclusive fitness of the organism's ancestors. Consider, for example, the fact that herring gull chicks instinctively peck at the red spots on the lower bills of adult gulls. This behavior has a function (procurement of food from adults) thanks to its contribution to inclusive fitness in the evolutionary history of gulls. Here it is natural selection which is serving to stabilize the task function.

One of the distinctive features of Shea's version of teleosemantics is its pluralism regarding the processes which can serve to stabilize outputs. (For another recent endorsement of pluralism, see Garson and Papineau forthcoming.) According to Shea, natural selection is just one of the processes suited to play a role in transforming outputs into task functions. He acknowledges two further forces, both operating at the level of *individuals* (as opposed to *populations*): feedback-based learning and contribution to persistence. Suppose an individual bee exhibits a preference for objects of a certain color and odor in its foraging behavior. In principle, the bee's behavior could count as goal-directed in virtue of its contribution to inclusive fitness in the evolutionary history of the species. Alternatively, imagine that in the first instance the bee's act of approaching that kind of object rather than available alternatives was more or less random and not yet goal-directed. And imagine that over time the output in question became robust thanks to positive reinforcement learning (with nectar as a reward). Here it is feedback-based learning that is serving to stabilize the task function. Finally, suppose that the behavior is a product of a random mutation in the individual bee and so not initially goal-directed. And suppose that the behavior in question makes a direct contribution to the bee's survival. We can grant that the bee's behavior eventually becomes goal-directed in virtue of its contribution to the persistence of the individual.

Shea's rich discussion of goal-directedness is, I think, a real highlight of the book. First, Shea does a nice job motivating an historical approach to function, arguing that non-historical views saddle us with an undesirable amount of indeterminacy regarding a system's functions and what counts as successful behavior. Second, he also offers a powerful case in favor of pluralism about stabilizing processes. He identifies important similarities between the three processes, but resists endorsing a single, overarching account on the grounds that a unified account would fall prey to counterexamples. Finally, Shea's treatment of the much-discussed case of swampman is easily the best I have encountered. Although Shea seems to be appropriately skeptical of the probative value of intuitions, he nonetheless offers an intuitively compelling response to familiar swampman-related worries about teleosemantics.

We now have the basic idea of a task function understood as an output produced robustly and stabilized in one of the three ways indicated. While a task function is a goal-directed output, a system need not deploy representations in performing a task function. Representations have a role in explaining performance of a task function only when the system implements an algorithm for performing the function, a sequence of operations on vehicles serving as stand-ins for distal states of affairs. This requirement on representational explanation has two components. First, the system must have internal vehicles bearing exploitable relations to distal states of the world relevant to the task in question. These vehicles are internal representations of the world. Second, the system must perform the task function *via* computations on these internal vehicles, computations which make use of the relations these vehicles bear to the world.

Shea further solidifies his commitment to pluralism about representation by distinguishing two kinds of exploitable relation: correlational information and structural correspondence. As with many aspects of the book, Shea's remarks on correlational information and structural correspondence offer refinements and elaborations on his earlier work on representation. Correlational information is the sort of information routinely exploited in associative learning. For example, a fruit's color can carry correlational information about how ripe it is. Shea's general approach to correlational information is a variation on views familiar from authors like Dretske (1981) and Skyrms (2010), who analyze the information carried by a signal in terms of changes

in probability. Meanwhile, structural correspondence is the kind of relation we exploit when we successfully navigate by deploying a 2D map. In successful navigation we exploit the way spatial relations between points on the map correspond to spatial relations between places in the world.

With these aspects of Shea's view in hand, we are now in a position to understand how the meaning or content of a representation is to be determined. For ease of exposition, I focus on signals in animal communication rather than signals in the brain. (Shea's view is intended to cover both types of representation. Note that the signals or representations involved in animal communication are external to the system with the task function, while the representations invoked in cognitive science are internal representations.) In the case of aposematic signaling, animals are exploiting correlational information carried by the signal (conspicuous coloration). But there are many things which correlate with this kind of signal. For example, the animal's coloration may correlate with aspects of its diet. How, then, do we settle which correlation captures the meaning of the signal? To answer this question, we ask a further question: Which correlation is causally responsible for the stabilization and robustness of the relevant task function? The task function relevant to aposematic signaling is the avoidance behavior on the part of the predators who receive the signal. We are asking what state of the world is causally responsible for the stabilization and robustness of this output, so we have to look to the historical forces which shaped the response of the receivers of the signal. One way aposematic signals work is through the stabilization process of feedback-based learning. An insectivore learns that the bright coloration of a wasp or bee correlates with unprofitability. Instrumental conditioning is the stabilizing force and unprofitability is the state of the world that is causally responsible for the stabilization in question. Another way aposematic signals get traction is through natural selection. Poisonous dart frogs can be so lethal that there is no opportunity for trial-and-error learning on the part of potential predators: the aversive behavior needs to be hard-wired. In this case the stabilizing process is natural selection and unprofitability is the state of the world causally responsible for the stability. But even before natural selection has had a chance to stabilize the signaling system, we can allow that an individual predator might have the task function in question. The aversive behavior might initially arise by random mutation in the individual and later prove to be crucial to the individual's survival. In this instance the stabilizing force would be contribution to the persistence of the individual and once again the causal force

is unprofitability. In all of these cases the signal has the same meaning: unprofitability for would-be predators.

We now have an overview of Shea's approach to representation. What does this approach tell us about the role of representation in making sense of behavior? Acknowledging representations allows us to explain two aspects of task performance: how the outcome is produced and why. The proximate explanation (how) of behavior driven by internal representations invokes processing on these representations. The system has internal vehicles bearing exploitable relations to task-relevant aspects of the system's environment, i.e. representations. The system implements an algorithm understood as a sequence of internal operations on these representations, the step-by-step transitions by which the system carries out the task function. Understanding how the system performs the task is partly a matter of grasping how this internal processing is making use of the exploitable relations to the world. In virtue of being keyed into distal states of the environment, internal processing explains how the system coordinates behavior with states of the world. Next consider the ultimate explanation (why) of a task function. Instead of looking to the step-by-step processing immediately preceding the output, we look to the historical processes which served to stabilize the output. Returning to our example of a predator's avoidance response to an aposematic signal, the ultimate explanation of the avoidance behavior is given by the representational content of the signal, namely, unprofitability. This ultimate form of explanation is important because it secures conditions of success and failure for a task function and allows us to explain these successes and failures in representational terms. For example, successes and failures of the predator's task function are routinely explicable in terms of whether the aposematic signal is honest or dishonest (accurate or inaccurate), respectively.

I remarked at the outset that Shea has offered the most compelling version of teleosemantics to date. In the remainder of this review I will suggest some potential advantages of Shea's approach over traditional teleosemantics. Along the way I will also indicate some lingering worries I have about his theory.

The most striking difference between Shea's view and standard teleosemantics is his commitment to pluralism about stabilizing processes. Dretske (1988) argued that representations do not explain behavior when they are stabilized by natural selection. The following example captures Dretske's worry. Suppose a predator becomes geographically isolated from honest aposematic signals, and the only remaining signals of the kind in its local environment are dishonest (i.e. instances of Batesian mimicry). The signal no longer carries correlational information about unprofitability, so it no longer has the meaning it used to have. If the predator's task function (avoidance behavior) has been hard-wired through the process of natural selection, the predator will continue on with the behavior regardless of the loss in meaning. Dretske infers that meaning is irrelevant to the predator's behavior on the grounds that the predator behaves the way it does whether or not meaning is present. He suggests that behavior shaped by feedback-based learning is importantly different because it is sensitive to changes in meaning—conditioned behaviors are routinely extinguished over time when they are not sufficiently reinforced. For this reason Dretske rejects the kind of pluralist view Shea endorses.

Shea mentions Dretske's worry about his pluralism, but he does not explicitly respond to Dretske's argument. Here I sketch a response to the argument which is in line with Shea's overall position. Dretske is right to note that there is an important difference between the two cases. With behavior stabilized through instrumental conditioning, we can typically alter the behavior by manipulating the correlational information carried by the conditioned stimulus. Meanwhile, behavior stabilized through natural selection is relatively inflexible, and intervening on the correlational information carried by the triggering stimulus does not affect the behavior. Nevertheless, the two forms of behavior do have something important in common: intervening on the strength of the correlational information carried by a signal affects the success rate of the behavior. All else equal, by increasing the ratio of honest signalers to dishonest ones in the local environment of the predator we can increase the success rate of the task function. Likewise, by decreasing the ratio of honest signalers to dishonest ones we can decrease the success rate. Meaning, then, is very much relevant to the success of the behavior: in the limiting case where correlational information is removed and the meaning has thereby changed, the behavior will no longer have the kind of success it previously enjoyed.

The foregoing remarks highlight an advantage of Shea's view that the principal explanatory role of representational content is that of making sense of success and failure in goal-directed behavior. But the more we emphasize this goal of explaining success and failure, the less obvious it becomes why the behavior to be explained must satisfy Shea's robustness requirement. Shea's introduction of this requirement is another way his view departs from standard teleosemantics. But whereas Shea's pluralism broadens the reach of the teleosemantic program, the robustness requirement has the opposite effect. It seems to exclude phenomena which teleosemantics is well equipped to accommodate. Suppose a predator's avoidance response to an aposematic signal is rigid (non-robust) in the way that a rifle's firing response is: a highly specific output is set off by a highly specific input. If our goal is to explain the success or failure of the response, we can do that even for highly rigid behavior. Avoidance responses to honest aposematic signals are successful whether or not the responses are rigid.

Shea might reply that the robustness requirement is important because it helps us address Ramsey's job description concern. We need to distinguish representations from things that are merely serving as links (causal mediators) in the chain extending from stimulus to response, and the robustness requirement is supposed to highlight a distinctive role for representations. A distinctive feature of representations is that they are invoked to explain behavior produced in response to various inputs and across various conditions. It is unclear, however, whether the robustness requirement really provides the sort of distinguishing mark of representations we are looking for. It would seem that the robustness requirement can be satisfied when a creature has a back-up system for executing some task. Suppose a nocturnal organism is typically guided by visual cues in executing a specific foraging task, but in circumstances of extreme darkness the behavior is guided by olfactory cues instead. The robustness requirement appears to be satisfied, but all we have here is a second path extending from the distal stimulus to the output.

Standard teleosemantics as captured by the work of Millikan assimilates internal representations to signals in animal communication, signals figuring in an evolutionarily stable signaling system which coordinates activity of the organism receiving the signals with the states of the world represented by the sender. Shea's approach differs not only in allowing for forms of stabilization beyond natural selection; it also departs from standard teleosemantics in

eschewing commitment to sender and receiver units. While Shea is sympathetic with the idea that representations serve to coordinate actions with world states, he thinks that singling out *the* consumer/receiver of a neural representation is not always possible when confronted with the complexities of neural processing in the brain (cf. Cao 2012, 2014).

One thing I like about Shea's emphasis on task functions rather than functions of senders and receivers is that it provides a nice framework for addressing generalist discriminatory powers. Sensory ecologists regard some discriminatory powers as specialized on the grounds that they evolved to facilitate a specific behavioral response. For example, according to one popular view about the emergence of trichromacy in primates (the frugivory hypothesis), it was specifically the act of selecting ripe fruit among green foliage which favored trichromacy. In this case teleosemantics offers a clear verdict regarding the content of the discriminatory state stabilized by the presence of ripe fruit: it represents ripe fruit. Other discriminatory task functions are generalist in nature. For example, sensory ecologists treat avian color vision as a generalist system because it is remarkably uniform across species differing considerably in their ecology and behavioral adaptations. The meaning of these all-purpose signals is much less clear within standard teleosemantics because it is far from obvious what is supposed to play the role of content-constituting consumer. Obviously, this worry does not arise for Shea, who rejects the need for content-constituting consumers. He accommodates multi-purpose discriminatory states by allowing that representation can emerge through the stabilization of multiple task functions.

While Shea's account seems better equipped than traditional teleosemantics to handle generalist discriminatory powers, it falls prey to a problem concerning specialized discriminatory powers. Suppose the frugivory hypothesis is correct. Shea seems to be committed to the consequence that metamers will be misrepresented as possessing the ecological feature in question (ripeness). This consequence is problematic because there are strong theoretical reasons for supposing that metamerism is a form of limitation rather than error or misrepresentation (Ganson 2018a).

One important kind of criticism of standard teleosemantics is that it is unable to accommodate systematic misrepresentation. Elsewhere I have discussed how Shea's view handles one

version of this worry (Ganson 2018b). Here I focus on different versions of the worry, ones not addressed by Shea himself. Mendelovici (2013) and McLaughlin (2016) raise doubts about whether teleosemantics can handle systematic misrepresentation under completely normal conditions. Mendelovici discusses the possibility that we always misrepresent objects as possessing colors in a colorless world. Shea can straightforwardly address this worry: representations are invoked to explain success and failure in performance of task functions and Mendelovici gives us no reason to think we systematically fail in the performance of color-related tasks (see Ganson 2018a). However, McLaughlin points to robust psychophysical evidence that errors in visual tasks targeting spatial properties like size and shape are the norm. Traditionally teleosemantics has analyzed misrepresentation in terms of malfunction, but talk of malfunction seems to make sense only against the backdrop of successful functioning. Accordingly, traditional teleosemantics is poorly situated to handle the pervasive forms of error to which McLaughlin calls attention. Meanwhile, Shea can insist that errors in the performance of this type of task function presuppose training on the psychophysical task itself, and training typically occurs through the positive reinforcement of responses which are successful (by objective measures). So failure in visual tasks targeting, say, relative size exists against the backdrop of the successful performance which figured in the stabilization of the task function.

In this example Shea will say that the property causing stabilization of the task function is an external feature of the world, namely, relative size. So, the property represented is an external feature of the world. More generally, Shea endorses an externalist approach to representational contents: contents are constituted by worldly causes of the stabilization and robustness of task functions. A potential problem for Shea's externalism is generated by the phenomenon of sensory exploitation, which occurs when males take advantage of pre-existing biases of females in their representations guiding mate choice. What causes stabilization of the signal is not something external like male fitness, as in the case of a Zahavian signal; rather, a pre-existing quirk of the female sensory system itself causes stabilization in the population. (One complication in this case is that the sender and receiver may have different interests. Shea *et al*. (2018) suggest that the very same signal may have different meanings for the sender and the receiver in a case of this sort.)

Shea calls his version of teleosemantics “varitel semantics.” Whether or not this label catches on, I expect his approach to meaning will become one of the canonical positions in the literature. Indeed, it is a safe bet that his book will prove to be one of the most influential philosophical works on representation since the publication of the books by Millikan and Dretske mentioned at the outset. It should be widely read and discussed by philosophers and scientists with an interest in foundational issues in the study of behavior.*

Todd Ganson
Oberlin College
tganson@oberlin.edu

*I am very grateful to Nick Shea for valuable written feedback in the process of writing this review.

REFERENCES

Burge, Tyler 2010, *Origins of Objectivity* (Oxford: Oxford University Press)

Cao, Rosa 2012, ‘Teleosemantic Approaches to Information in the Brain’, in *Biology & Philosophy* 27: 49-71

Cao, Rosa 2014, ‘Signaling in the Brain’, in *Philosophy of Science* 81: 891-901

Dretske, Fred 1981, *Knowledge & the Flow of Information* (Cambridge, MA: The MIT Press)

Dretske, Fred 1988, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: The MIT Press)

Ganson, Todd 2018a, ‘Sensory Malfunctions, Limitations, and Trade-Offs’, in *Synthese* 195: 1705-1713

Ganson, Todd 2018b, ‘The Senses as Signalling Systems’ in *Australasian Journal of Philosophy* 96: 519-531

Garson, Justin, and Papineau, David forthcoming, ‘Teleosemantics, Selection, and Novel Contents’ in *Biology & Philosophy*

McLaughlin, Brain 2016, ‘The Skewed View from Here: Normal Geometrical Misperception’ in *Philosophical Topics* 44: 231-299

Mendelovici, Angela 2013, ‘Reliable Misrepresentation and Tracking Theories of Mental Representation’ in *Philosophical Studies* 165: 421-443

Millikan, Ruth 1984, *Language, Thought, and Other Biological Categories* (Cambridge, MA: The MIT Press)

Ramsey, William 2007, *Representation Reconsidered* (Cambridge: Cambridge University Press)

Schulte, Peter 2015, ‘Perceptual Representations: A Teleosemantic Answer to the Breadth-of-Application Problem’ in *Biology & Philosophy* 30: 119-136

Shea, Nicholas, Godfrey-Smith, Peter, and Cao, Rosa 2018, ‘Content in Simple Signalling Systems’, in *British Journal for the Philosophy of Science* 69: 1009-1035

Skyrms, Brian 2010, *Signals: Evolution, Learning, & Information* (Oxford: Oxford University Press)

Sterelny, Kim 1995, ‘Basic Minds’ in *Philosophical Perspectives* 9: 251-270