Representation in Cognitive Science: Content without Function[1]

Robert D. Rupert
University of Colorado, Boulder
robert.rupert@colorado.edu
Feb. 21, 2021

For a book symposium on Nicholas Shea's *Representations in Cognitive Science*, in *Studies in the History and Philosophy of Science*

I. Shea's Project

Cognitive science is rife with talk of representations. The brain builds navigation-guiding spatial representations from the activity of place-specific cells in hippocampus (Ekstrom et al. 2003, Derdikman and Moser 2010). Computational models of visual processing include edge-detectors, which contribute to object-categorization via matching of portions of a constructed image to stored representations of categories (Marr 1982), which matching then leads to category-appropriate behavior. A given level of activation in midbrain dopamine neurons represents the difference between the predicted payoff of an action taken and the actual payoff of that action, with the effect, in some circumstances, of decreasing the probability that such behavior will be produced in the future (Schultz 1998). The development of a certain theory-like structure confers upon the child the capacity to represent, and thus categorize and reason about, mammals (Carey 1985). A given symbol string represents the syntactic structure of a heard sentence, and thereby facilitates the production of a content-relevant verbal response (Chomsky 1980). And so on. Sometimes cognitive scientists talk of concepts, symbols, or knowledge structures, rather than of representations; but despite what can be important differences (in complexity of the units involved, for example), common threads run through the discussion of all such structures. They are *about* the feature (or property, or kind, or individual – take this as read hereafter) represented;

[1] An earlier version of this material was presented at the workshop *Representation in Cognitive Science*, held at Ruhr University, Bochum, Feb. 3–4, 2020. My thanks to Tobias Schlicht and everyone who helped to organize the event and to the audience members, including Nick Shea, for fruitful discussion.

they stand in for it, depict it, track it, correspond to it, or are accurate when applied to it. Moreover, the structures in question either directly or indirectly help to produce behavior responsive to or distinctively directed toward what they are about (or depict, etc.).

Philosophers naturally wonder, then, what makes a given structure a representation of *this* particular feature, rather than some other. In his brilliant book, *Representation in Cognitive* Science, Nicholas Shea makes substantive progress on these questions, and he does so by attending, in a careful, empirically informed way, to the role that representations play in cognitive-scientific modeling. The result is what Shea calls 'Varitel Semantics'. According to Varitel Semantics, cognitive agents execute task functions; they successfully perform certain tasks consistently and robustly, that is, across a significantly varying range of contexts and input values. And, in cases in which representations are in play, agents' doing so is best explained by their implementation of algorithms (75)[2] operating over nonsemantically characterized vehicles.[3] Suppressing some details, a vehicle's representational content is whatever correlation or structural correspondence figures into the best causal explanation of the stabilization of a robust task function (83, 200), that is, the correlation or correspondence that accounts for the successful execution of task function (and the algorithm that supports it) (56, 69, 167, 199); where representations appear, stabilization and robustness are causally explained partly by the correlation or correspondence between vehicles over which the algorithm in question operates and existing features in the world (preferably objectively existing ones – 90, 122, 155, 158).[4] Three kinds of causal process can explain the stabilization of task functions: evolutionary

[2] All page references are to *Representation in Cognitive Science*, unless otherwise indicated.
[3] That is, symbols or other units that are characterized in terms other than semantic or representational terms, similar to characterizing a word by the letters that are strung together to make it up.
[4] *Cf*. the discussion of the natural-kinds-only assumption in Rupert (1999, 344–348)

processes, individual learning, or the contribution of the performance of the task function to ongoing persistence.[5]

As a brief illustration,[6] consider subjects who are trained to distinguish two kinds of images, As and Bs, based on the visual features of the stimuli. In doing so, subjects' cognitive or neural systems form "hidden" units (in contrast to input units that detect visual features) the activation of which corresponds systematically to the appearance of relevant stimuli, $a_1...a_n$, $b_1...b_n$. Such hidden units cluster into two groups, each of which causes one of two responses, 'A' or 'B' (which then, in a whole-agent system, might cause a further action, of, e.g., putting a_4 into bin A). The best explanation of how the subject acquired the ability to perform this task function appeals to the correlation between units over which the categorization algorithm operates and the corresponding individuals in the environment. Because that correlation was in place during training, subjects frequently succeeded in the categorization task, which success led to reward and thus to the entrenchment of the categorization algorithm in question. Thus, the units in question represent the corresponding individuals $a_1...a_n$, $b_1...b_n$.

II. Content without Function or Success

I applaud Shea's methodology, particularly the extent to which he treats the question of representational content as an issue in philosophy of science. I'm concerned, however, that Shea reads into cognitive science too deep a normative dimension,[7] despite Shea's attempts to distance

[5] Shea also recognizes the possibility that a system can perform a function because it's been assigned that function by an external system, a human agent, for example. Shea mostly sidelines these cases, however, because they do not shed light on the ultimate determinants of content; they do not help us to understand how representational content appears in the first instance, that is, how underived content can appear in a natural world; for such cases presupposes that the agent's intentions already have content, which content effects the assignment of function (9, 65).

Note that, throughout, I focus on learning and evolutionary selection as sources of stabilization, in contrast to persistence-related considerations. The role of persistence-related processes strikes me as provocative and worthy of further exploration but also as raising a host of complications and unaddressed questions treatment of which is prevented by limitations of space.

[6] This is a simplified version of the example discussed at (91–93).

[7] In the philosopher's sense of 'normative' – having to do with good, bad, right, wrong, ought, and should.

himself from the normative perspective (65); for Shea's project rests fundamentally on the distinction between successful and unsuccessful behavior. On Shea's view, the very target of representational explanation in cognitive science "is a pattern of successful behavior of a system in its environment" (22). In fact, for Shea, if a creature has not been in existence long enough to have had the computing of a definite function stabilized for robust application, representational content does not (yet) exist in that creature; if nothing counts determinately as behavioral success – behavior that has been stabilized because it is successful – there are no representations, on Shea's view (69, 167).

This seems to me to misrepresent the structure of cognitive science. The target of its explanations is not, inherently, a matter of behavioral success and failure. Rather, cognitive science is, first and foremost, an attempt to model processes that produce, in human behavior, deviations from the null hypothesis. More generally speaking, humans exhibit a range of unexpected, unusual, or distinctive forms of behavior – from the products of art to architecture to game play to social interaction to theorem proving – and cognitive science has bet that a dedicated, relatively unified scientific endeavor can explain them.[8] For example, the null hypothesis predicts that the text on a page that a subject has been exposed to visually will not correlate with what the subject produces when asked, at a later time, to provide written answers to questions about that page of text;[9] moreover, the null hypothesis predicts that the subject's

[8] "What questions should a theory of problem solving answer?" Newell, Shaw, and Simon asked at the dawn of cognitive science; and responding to their own question, "First, it should predict the performance of a problem solver handling specified tasks" (Newell, Shaw, and Simon 1958, 151).

[9] In a rigorously worked-out case, such questions themselves would be scored by an independent measure, so as to qualify for the status of being, as we would normally say, "about the original text." When designing and running experiments, researchers sometimes take for granted their ability to recognize such relations as "being relevant questions about text x," but frequently efforts are made to ground such assumptions, by, e.g., asking subjects to score questions or by using analytic methods such as correlation-based tools (deploying latent semantic analysis, for instance). Ultimately, though, even such approaches must take something tentatively for granted, for example, taking a corpus as given. But, cognitive science faces no special problem here that reliance on normativity might solve. Every science faces the foundational problem of grouping together phenomena or data sets for study, judging

being asked to produce written answers does not modulate the strength of the (null) correlation between the text the subject was exposed to and the written text the subject produces after exposure. But, in fact, human populations deviate significantly from the null hypothesis in both respects.[10] If it is not clear, I've described performance on a reading comprehension exam by those who, as we would say informally, speak the same language as that in which the text is written and is spoken by those administering the test. In the everyday human case, in literate societies, we observe significant deviations from the null hypothesis: the subject population of interest – speakers of the same language – produces answers systematically correlated with aspects of the written text; and they do so at an even higher rate when they've been asked specifically to write about that text by experimenters with whom they share a language. The job of cognitive science is to model the mechanisms and processes that mediate such correlations.[11] This vision of cognitive science does not essentially involve success or failure – despite the fact that more than one parent has responded to a child's school results in such terms.

Thus, although it's natural to consider some patterns of behavior as successful or unsuccessful, the doing of cognitive science – model building, model selection – does not take the explanation of successful behavior as its central goal. This is no surprise; it is the pattern seen throughout the sciences. Humans place various natural phenomena in normative categories. Hurricanes and turbulence are bad, while bumper crops and sunny days are good; but the *explananda* of meteorology, fluid dynamics, and agricultural science are not success and failure,

---

them to be part of a family of related phenomena and guessing that the treatment of them as such will prove fruitful. If judgments about which texts go together in a corpus or judgements concerning the measure to use to determine that subjects speak the same language as the experimenters do not lead to fruitful experimental results, then it's back to the drawing board.

[10] Which populations? Those who speak and read the language in which the experiment is conducted. Perhaps that's identified by clustering together those whose average length of verbal interaction with each other deviates significantly from baseline, which measure might operationalize "speaking the same language," depending on the context of the interaction.

[11] C*f*. the methodological framework for social psychology articulated by de Houwer, Gawronski, and Barnes-Holmes (2013).

qua success and failure. The distinction between *any* two kinds of outcome can be of interest to a science and be the target of a modeling enterprise. It's of no deep consequence that, in some cases, the difference between the two outcomes of interest have positive and negative valence attached to them or attitudes directed toward them.[12]

Consider, too, the extent to which cognitive science examines what are thought of as dysfunctional or pathological cases, to ask, for instance, what accounts for the difference between the actions typical of someone with delusions of grandeur and the actions typical of someone with delusions of being dead (Cotard's syndrome) or to ask why subjects develop one delusion rather than the other. Such explanations are likely to be deeply representational, but talk of success and failure gets little grip. And less extreme cases abound: surely cognitive science would like to detail the mechanisms that produce such arational actions as throwing a nonfunctional can-opener across the room in frustration or ruffling the hair of a loved one when walking past (Hursthouse 1991), as well as such results as subjects' unwillingness to eat feces-shaped pieces of fudge, even when they know full well that what they're faced with is fudge (Gendler 2008). Yet, talk of success and failure makes little sense in these cases. And the list is, indeed, quite extensive: schizophrenic reports of auditory hallucinations, perseveration, varying patterns of resistance to evidence, construction of individual identity in a social context,

[12] This might seem to be too fine a philosophical point, but in this context, it's worth emphasizing. We do often categorize behavioral responses as successes or failures, and we are, in fact, out to explain those responses; so, in an indirect sense, we do mean to explain successes and failures. But, cognitive science is not out to explain them *as* successes or failures or to explain their status of 'being a success' or 'being a failure'. Cognitive science is out to explain differing patterns of output – the button-pressing did or did not correspond with the appearance of the stimuli in question – and often such patterns can be categorized in success-oriented terms, but that success-based categorization is, generally speaking, incidental to the logic of scientific modeling. Shea's project requires that being a success or a failure play or more substantive role than that – of being the very thing that the modeling effort aims at – which is what I take exception to. (Of course, we might shift gears and attempt to model the psychological processes by which subjects produce the judgments that a response is a success or a failure. Why does the parent think the child's 'D' on a reading comprehension exam is a bad outcome? Why, more generally, do subjects deviate from the null hypothesis in their labeling of behavioral patterns as successes or failures? But that way of pursuing questions about behavioral success and failure, as an investigation merely of patterns of labeling actions, reinforces the conception of cognitive science I've been pressing.)

addiction, mimicking of speech patterns and bodily movements of one's interlocutor, priming-related behavior, reports of mind-wandering, novelty-seeking behavior, and risk-taking behavior. In what sense is it a behavioral success or a failure that, after having been exposed subliminally to 'stamp', the subject is more likely to generate the word 'lamp' when asked to write down names of pieces of furniture, the sort of result one observes in a typical priming-based experiment?

It's worth considering possible ways to insulate Shea's project from these concerns. Here are two. First, many forms of behavior that seem irrational or pathological can seem objectively successful (or unsuccessful) when relativized to the subject's existing goals. On one reading of this response – one that relativizes success to the subject's explicit goals – it violates Shea's naturalistic ground rules (9, 65); it requires that the content of the subject's goals already be in place.[13] Another reading of this proposal appeals to biological imperatives, but this path offers scant sustenance. Cognitive scientists should want to identify the mechanisms that produce paintings, poetry, and pontification, but there's no straightforward, nonrepresentational route from the biological goals of, say, locating food and shelter to, for instance, the motivation for the doodlings on the cover of my notebook. And, to turn to cases of much greater gravity, I doubt that such a route, if it were found, would account for the behavior of individuals who are suicidal or who, in the case of eating disorders, refuse food. One would think, however, that in all of

[13] Shea's own approach to the typing of vehicles and of syntactically individuated units seems to close off a further response, that of appealing to behavioral success and failure to fix content in a range of base cases, and then treating content, in the non-success-involving cases, as parasitic on the base cases. On Shea's view, however, semantic content follows sameness of syntactic type, not sameness of vehicle type, yet sameness of syntactic type requires that tokens of the same type are "processed the same way by the system" (39). The requirement of sameness in processing seems to limit severely the range of cases in which Shea might say that a unit appearing in pathological, arational, etc. processing has content on the grounds that the unit also participates in distinct, success-grounded processing; the processing of the unit in the base case as contrasted with the other cases would seem to differ to too great an extent to preserve syntactic type. This strikes me as a general potential problem with Shea's approach to unit-typing, given how widespread I suspect it is for a unit's content to be preserved across processing contexts (despite the likelihood of contextual shifting of the content attached to a given physically individuated unit – Anderson 2014).

these cases, the cognitive-scientific explanation of the relevant forms of behavior will appeal to representations.

A second possible defense of Shea's conception of cognitive science appeals to the determinacy of the functions being computed and its relation to success-conditions. To be clear, Shea appeals explicitly and repeatedly to behavioral success and failure, so the concerns expressed above strike me as fair. Nevertheless, Shea's talk of algorithmically computed functions sometimes sounds more narrowly focused than any claim about the organism's behavioral output; he's concerned with the stabilization of a disposition to compute a certain function (rather than some other one), and perhaps it goes without saying that the identity of a function – in the mathematical sense relevant to the theory of computation – entails determinately correct and incorrect outputs. So, Shea might limit his focus to cases involving the computation of a specific function, perhaps ignoring the distance between a function (in the narrow sense) computed and the whole agent's behavior. In this way, Shea might identify success conditions connected to many steps in the production of even the most pathological or biologically counter-productive behavior.

This approach may hold promise, but undoubtedly there be monsters. To the extent that human behavior is the computing of functions, it is a mess of noise and imprecision. Even when bits of the brain compute functions, the process is often sloppy and imprecise, and which functions are being computed is shifting and frequently assembled on the fly (Anderson 2014). The functions that seem most likely to be stable and precise are the most narrowly defined (e.g., "compute retinal disparity"); but, at least if we focus on more complex and interesting forms of behavior – the results of so-called higher cognition – these functions are unlikely to issue directly in behavior; rather, their effects will be combined with and modulated by architectural facts,

further operations, mechanisms computing other functions, and more, all of which co-contribute to the production of the behavior that a narrowly defined function helps to produce. As often as not, orderly behavior emerges from the brain by hook or crook, by passing control among a shifting set of dominant networks and subnetworks, each of which involves various areas and the computing of many functions at simultaneously – bleeding into one another, interfering with one another, competing with one another. Reinforcement algorithms (and their evolutionary analogues) might solve credit-assignment problems well enough, allowing for the application of Shea's framework to whatever determinate functions are being computed in the production of behavior, perhaps thereby allowing for the grounding of all representational content in the computing of such functions, but this is an ambitious hope.[14] [15]

To be fair, much of Shea's framework might prove exceptionally useful, even if he abandons his normative conception of the *explananda* of cognitive science. The workings of complex, kludgey processes consistently and robustly produce human behavior that deviates from the null hypothesis. And, it's not implausible that causal explanations of the (relative)

[14] Shea might lean on the competence-performance distinction here to shore up claims to determinacy, but this kind of response has its limits. The messiness of the human brain's functioning – the extent to which approximations of function emerge from complex interactions of noisy, imprecise, biology-corrupted computations – belie claims to determinate competence, fixed by learning or evolutionary selection. Compare: I might claim that I have the competence to play tennis as well as Roger Federer – a competence I simply never perform in keeping with – but someone who looks inside me at the motor and muscular mechanisms that generate my on-court behavior will rightly see through my claim to competence.

[15] Many influential cognitive scientists might seem to be in Shea's corner: David Marr, David Robinson, John Anderson, and Tom Griffiths all have, at one point or another, emphasized ideal rationality or abstract function over messy details. But, upon closer look, such authors often pitch the computational level, adaptive rationality, and Bayesianism about the brain as part of what used to be called "the context of discovery"; thinking in terms of optimal behavior is an efficient way of discovering models of the (typically messy) mechanisms that produce actual human behavior, which is normally only a very limited approximation to optimality. For Marr, the computational level clearly plays a guiding, epistemic role (1982, 27). We don't stand a chance, he seems to think, of figuring out what the brain is doing if we don't ask what input-output functions would be useful for it to compute. For Griffiths, it's matter of exploring the space of models efficiently; we could proceed bottom-up, from neural mechanisms to the construction of models of performance, but that approach is likely to lead to a long string of dead ends, whereas focus on optimality, as a working assumption, will likely lead more quickly to successful models of (suboptimal) performance (e.g., Griffiths et al. 2010, 363).

stabilization of that robustness could ground representational content in at least roughly the way Shea describes.[16]

III. Pluralism, Multidimensional Analysis, and the Personal Level

I suggest recasting the discussion of mental representation in the language of statistical modeling. We should define historical, possibly content-fixing relations – teleological or learning-based – as variables in a collection of quantities the dependence relations among which might or might not be correlated in significant ways. The strength, over some definite period of time (e.g., over the subject's past), of the co-occurrence of the activation of the unit in question and the appearance of some possibly represented feature might itself be represented as a variable. We can then ask to what extent that quantity accounts for variance in other quantities measured behaviorally; [17] to put matters more crudely, we can ask such questions as, "Does the strength of the co-variation between Pat's appearance in the room and the activation of the unit in question, as determined by the subject's history, account for variance in the probability that the subject greets Pat when Pat is encountered at testing time?" This way of approaching representational content accommodates a multivariate approach to content determination, allowing us to see how different kinds of processes or relations – possibly including stabilization by learning, past causal interaction, evolutionary selection, and others – might contribute, to differing degrees, to the production of the subject's behavior, with different factors accounting for different amounts of the subject's deviation from the null hypothesis.[18]

---

[16] I myself attempted to work out a story in this vicinity (Rupert 1998, 1999, 2001), one focusing heavily on developmental processes; for an update of sorts, see Rupert (2018*a*). For a project embracing mental representation without rules or tidy computing, see Horgan and Tienson (1996).

[17] Shea's discussion of the evidential test (89–91) suggests this kind of approach. Note that the quantities in question can include the extent to which a correlation has contributed to stabilization in the subject's use of a given algorithm.

[18] *Cf.* Shea's comments about pluralism (66).

Such an approach might naturally allow Shea to abandon what strikes me as his unmotivated deference to the so-called personal-subpersonal distinction (8, 26, 42, 134, 162), which, so far as I can tell, plays no role in cognitive-scientific modeling (Rupert 2018*b*). Consider one of the four cases that Shea treats as a personal-level form of behavior and thus treats as off limits in the context of his project: "explaining to others what we believe and why we act as we do" (26). Clearly cognitive science should want to model such behavior; it should not be set aside when developing a naturalistic account of representational content informed by the role of representation in cognitive science. Of course, one might choose to call states that produce such behavior "personal-level states," but the implication concerning levels seems downright misleading (Rey 2001, Drayson 2014). Such states contribute, as nodes in a network model or in some other form of statistical model, to the production of the relevant behavior, right alongside states that Shea would categorize as subpersonal. The result of banishing the personal-subpersonal distinction might be greater fidelity to cognitive science, perhaps via the development of multivariate models that include explicit attitudes (Perugini 2005) and conscious states in the same analytic framework as neural bits that, for instance, compute retinal disparity. This would allow Shea to pursue a more inclusive approach to cognitive science, one that extends to all aspects of it – including accounts of the forms of behavior associated with the phenomena he brackets as personal level – when developing and applying a naturalistic semantics for mental representations.

Works cited

Anderson, M. L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain.* Cambridge, MA: MIT Press.

Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chomsky, N. 1980. *Rules and Representations* (New York: Columbia University Press).

De Houwer, J., B. Gawronski, and D. Barnes-Holmes. 2013. "A Functional-Cognitive Framework for Attitude Research," *European Review of Social Psychology* 24, 1: 252–287.

Derdikman, D., and E. I. Moser. 2010. "A Manifold of Spatial Maps in the Brain," *Trends in Cognitive Science* 14, 12: 561–569.

Drayson, Z. 2014. "The Personal/Subpersonal Distinction," *Philosophy Compass* 9, 5: 338–346.

Ekstrom, A. D., M. J. Kahana, J. B. Caplan, T. A. Fields, E. A. Isham, E. L. Newman, & I. Fried. 2003. "Cellular Networks Underlying Human Spatial Navigation," *Nature* 425: 184–187.

Gendler, T. 2008. "Alief and Belief," *Journal of Philosophy* 105: 634–663.

Griffiths, T. L., N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. 2010. "Probabilistic Models of Cognition: Exploring Representations and Inductive Biases," *Trends in Cognitive Sciences* 14: 357–364.

Horgan, T., and J. Tienson. 1996. *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.

Hursthouse, R. 1991. "Arational Actions," *Journal of Philosophy* 88, 2: 57–68.

Marr, D. 1982. *Vision*. New York, NY: W. H. Freeman and Company.

Newell, A., J. C. Shaw, and H. A. Simon. 1958. "Elements of a Theory of Human Problem Solving." *Psychological Review* 65 (3): 151–166.

Perugini, M. 2005. "Predictive Models of Implicit and Explicit Attitudes," *British Journal of Social Psychology* 44: 29–45.

Rey, G. 2001. "Physicalism and Psychology: A Plea for a Substantive Philosophy of Mind." In C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents* (Cambridge: Cambridge University Press), pp. 99–128.

Rupert, R. D. 1998. "On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought," *Synthese* 117: 95–131.

Rupert, R. D. 1999. "The Best Test Theory of Extension: First Principle(s)," *Mind & Language* 14: 321–355.

Rupert, R. D. 2001. "Coining Terms in the Language of Thought: Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory of Mental Content," *Journal of Philosophy* 98: 499–530.

Rupert, R. D. 2018*a*. “Representation and Mental Representation,” *Philosophical Explorations* 21, 2: 204-225.

Rupert, R. D. 2018*b*. “The Self in the Age of Cognitive Science: Decoupling the Self from the Personal Level,” *Philosophic Exchange* 47: 1–36.

Schultz, W. 1998. “Predictive Reward Signal of Dopamine Neurons,” *Journal of Neurophysiology* 80, 1: 1–27.

Shea, N. 2018. *Representation in Cognitive Science*. Oxford: Oxford University Press.