Forthcoming in *Studies in History and Philosophy of Science*
Please cite published version, doi.org/10.1016/j.shpsa.2021.05.012

***Studies in HPS* Book Forum**

Representation in Cognitive Science: Author's Reply

Nicholas Shea

It is a rare privilege to have such eminent and insightful reviewers. Their kind words about the book are much appreciated – perhaps more than they realise. And I'm grateful to all three for having read the book so constructively. Each has given me several things to think about. In the space available here I will focus on the objections that seem most critical. Robert Rupert argues that I rely on an overly narrow understanding of what the cognitive sciences explain (§1). Elisabeth Camp presses me on what precisely it takes to qualify as a structural representation and raises questions about holism (§2). John Krakauer makes a fundamental objection to positing representations when they seem not to be needed to explain behaviour (§3). Rupert has also provided a useful introduction to the book, so I will jump straight in with my replies.

(1) Rupert: Representations Explain More Than Behavioural Success

I suspect that Rupert is right that much work in cognitive science does not explicitly set out to explain success and failure. However, explaining behaviour undoubtedly is a central concern. I claim that representational content is central to explaining behaviour. Characteristic of content is the possibility of misrepresentation. What does misrepresentation explain? Answer: something going wrong in some way downstream, often in behaviour. The other side of the coin – only slightly less obvious – is that representing correctly (often just 'representing', unmarked) can explain cases where there is no failure in the behaviour that eventuates, when things go well in some sense, or proceed in a regular way.

By no means all outputs are the proper target for this kind of explanation: for explaining success in terms of correct representation. Rupert lists many examples. What identifies the class of behaviours that can count as successes? That is a meta-level question, a question for the theorist, on which I don't take cognitive scientists to be authoritative. There are hints in the literature, but I take my task primarily the be one of systematising and theorising about standard explanatory practice (which may on occasion involve re-classifying some of the cases). My hypothesis is that, at least in many central cases in cognitive science, what picks out the outputs that count as behavioural successes (i.e. candidates to be explained by correct representation) is that they are task functions. I also think there is a deep explanation of why that is so. There is a cluster of features in nature that underpins the wide applicability of the representationalist form of explanation. Briefly, stabilising processes select for robustness in outcomes, and a good way of achieving that is to rely on exploitable relations and internal computations.

Rupert offers an alternative hypothesis, namely that the cases to be explained are those that differ from the null hypothesis. However, even leaving aside the question of what should count as the null hypothesis, this seems to me to switch to a wider question: what forms of behaviour does cognitive science set out to explain? I am asking which explananda should count as behavioural successes whereas Rupert is asking which explananda fall within the ambit of cognitive science. I agree with his catalogue of examples as an answer to the latter question. Indeed, many examples illustrate the other half of the characteristic explanatory grammar I rely on. They are cases where misrepresentation explains failure. For example, in cases of delusion, part of why patients suffer dysfunction in their lives is because they radically misrepresent the way the world is. In other cases, e.g. with lesions or priming, the behaviour is explained in terms of properties of representational vehicles, not their semantic contents. These cases are not just consistent with my view – I rely on them as evidence for internal representation. They are also part of what gives representationalism its distinctive explanatory purchase ((Shea, 2018), §2.5). My claim is that the explanatory grammar – representing and misrepresenting explaining behavioural success and failure – helps us to give a theory of what representations are and how their contents are fixed. This opens the way to understanding a much wider class of cases: the many and diverse ways in which representations and their contents are appealed to in cognitive science.[1]

Rupert's remarks about hurricanes and so on suggest that he may have taken my success / failure to be more normative than it is, to carry a positive or negative evaluation. Whether a behaviour and its outcomes are good or bad is, however, a separate question. Successful behaviour produced by correct representations can count as good (catching a child falling off a climbing frame) or bad (stabbing someone in the heart). The same goes for behavioural failures. I agree that the explananda of meteorology, fluid dynamics, and many other sciences are not successes and failures (although they too can be good or bad). This difference is part of what makes representational explanation a special case – and a puzzling one. By having that difference in its sights, I hope that my theory can go some way towards dissolving the puzzlement. Attempts to explain content in terms of nomic probabilistic generalisations, in the way Rupert suggests, similarly face the challenge of identifying what is special about the representational case (for one promising attempt see (Usher, 2001)). Covariance of features in a multidimensional data set seems not to be proprietary to cognitive science, but central to causal explanation throughout the sciences. We can hope that a good theory of content will show us what is distinctive about representation.

(2) Camp: Structural Representation and Holism

Elisabeth Camp deftly summarises what I say about different kinds of representationally significant structure. She then pushes hard on these distinctions. Rightly so. First off we have what Peter Godfrey-Smith has called an 'organized sign

[1] My focus is just on part of cognitive science. I set aside beliefs, desires and conscious states. Although these are also studied by cognitive science, I am silent about them in the book – not because I have a positive argument that content is constituted differently there, but because giving an account of content even for the simpler cases is hard enough.

system' (Godfrey-Smith, 2017). When a system of representations displays 'organization' in this sense, there is a systematic relation between a vehicle property and content. For example, with an analogue magnitude representation, activation might covary linearly with the quantity represented. I say that content may arise from correlations carried at the level of these types ('exploitable correlation carried by a range of states', p. 78). Downstream processes respond to instances of the type in a systematic way, and that is in a broad sense useful. But that is not yet a matter of the relation between two or more representations itself becoming a content-bearing entity.

The hippocampal cognitive map is supposed to illustrate what more is needed to qualify as a structural representation. In a s*tructural representation,* a relation on representational vehicles represents a relation on the entities represented by those vehicles (p. 118). Camp shows that we need more detail before we can conclude that hippocampal place cells implement a structural representation or amount to a map. She ingeniously distinguishes between three ways place cells could be deployed, neatly illustrating the key distinctions. Maps have a holistic representational structure and only the third case is fully holistic. However, a structural representation need not be holistic. The first implementation is neither holistic nor a case of structural representation. In Camp-System2 a structural representation is at work. It is nevertheless much less holistic than Camp-System3, which deploys a holistic structural representation.

Camp-System1 records a list of cell-pairs between which the rat has transitioned. Each entry in the list is like a little sentence (e.g. $L_1 \rightarrow L_2$). Co-activation may have been the basis on which new pairs were added to the list, but the list itself is a standing record of accessibility relations. This implementation would not vindicate my claim that the co-activation relation represents the relation of spatial proximity. And it is consistent with much of the data I cite about place cell replay. Camp-System2 does make use of the co-activation relation when it is calculating potential routes. It relies on the existence of a co-activation relation, and the time taken for that transition, to track how close together the corresponding locations are. It uses that relation in a uniform way across a number of co-activation sequences. That amounts to exploiting a correspondence between the relation of co-activation on place cells and the relation of proximity on places. My account therefore implies that this is a case of structural representation. However, I agree with Camp that this would not vindicate the label 'cognitive map'. Only Camp-System3 has a functional implementation of the kind of holism displayed by familiar cartographic maps.

These cases neatly show that we need to take account of seemingly fine-grained differences in the way representational vehicles are organised – rather than any simple dichotomy – to understand how representations work in practice. I do think, however, that there is one sense in which Camp-System2 is more holistic than Camp-System1. In Camp-System1 the base representations are represented pairs of locations ($L_1 \rightarrow L_2$). In Camp-System2, the base representations are chains of co-activation (from L_1 to L_2 to L_4 to L_6). A token in the chain – an instance of place cell firing – stands in relations of co-activation to all the other tokens in the chain, both its neighbours and beyond. Putting it in the chain puts it into lots of relations at once. This is the same principle as with a map, albeit on a very limited scale. Since the different chains are each tokened separately, this falls far short of the more widespread holism of Camp-System3. In Camp-System3, the pattern of spreading activation is one large structural representation, holistically representing spatial relations across much of the rat's environment.

This connects with the issue of how new representations arise. Having a regular way to coin new representations is definitely another 'good-making' feature of a representational system. Camp shows that there are a range of interesting – and interestingly different – ways of doing that. We could start with location sensitivity driven by visual features and then learn the relations. Or we could start with an array of relations and learn what those locations look like, connecting up the nodes to visual input in an appropriate way. New evidence suggests that the answer is probably: both at once. As well as place cells, there is a whole armoury of other specialised components in the medial temporal lobe involved in spatial navigation: grid cells, head direction cells, landmark cells, object-vector cells and border cells. Recent evidence suggests that many of these are also involved in 'navigating' around other kinds of relational structure (Whittington et al., 2020). The model those researchers use to account for their data, the 'Tolman-Eichenbaum Machine', actually learns both at the same time: the sensory sensitivity and the relational structure.

I also agree with Camp that, in principle, such a system could come to represent higher-order relations over the first order relations represented. One note of caution however. It could be that higher-order relations are taken for granted in the way computations rely on relations between place cells. That would just mean that higher-order relations are represented implicitly. We need to distinguish implicit representation from explicit representation (Shea, 2015). A representation of a higher-order relation is explicit when there is a separate vehicle property that makes a difference to downstream computation and is relied on for its correlation or correspondence to the higher-order relation. So explicitly representing a higher-order relation is an additional achievement. Camp recognises this and astutely suggests that we should look for system-wide properties which play the relevant computational role. Rather like the way ensemble properties are extracted from a perceptual array, system-level properties of a cognitive map could be tracked directly. For example, the sparsity of a place cell array could be exploited for carrying information about how densely or sparsely resources are distributed in the corresponding domain. It would then amount to an explicit second-order representation.

Finally, I should note that I wholeheartedly agree with Camp that, as we look carefully at the underlying mechanisms, we're going to have to give up a neat dichotomy between 'perceptual, stimulus-dependent, unstructured, iconic representations' and 'conceptual, stimulus-independent, systematic, propositional ones'. In particular, to understand personal-level thought and planning, we'll have to mix and match. In thinking about concepts, I have found Camp's own work on 'characterisations' especially useful here (Camp, 2015).

(3) Krakauer: Models and Decoupling

John Krakauer's critique raises an important question about the whole project, indeed about the wider programme of which my book forms part. His broad focus nicely complements the other two reviews. Krakauer doubts that my account provides a sufficient condition for representation. It applies to systems that are too unsophisticated to harbour real representations. Having representations at least requires working with some kind of decoupled model. My account applies to systems that are model-free and lack decoupling, as well as to their more sophisticated cousins. A notion of representation that can apply to such simple systems is too weak to be the basis of a representational theory of mind.

Cognitive science has a long history to trying to identify forms of behaviour for which representations are essential. That was useful to address scepticism about the very existence of representations (as content-carrying particulars). But Krakauer assumes something stronger: that an adequate notion of representation should be applicable only when behaviour is at least that complex. That seems to me to confuse an epistemic question – how to tell if a system has representations – with a metaphysical question – what makes it the case that a system has representations. One aim of my account is to sharply distinguish the two. We should not assume that there is no representation when there is 'no need for representation' to explain behaviour.

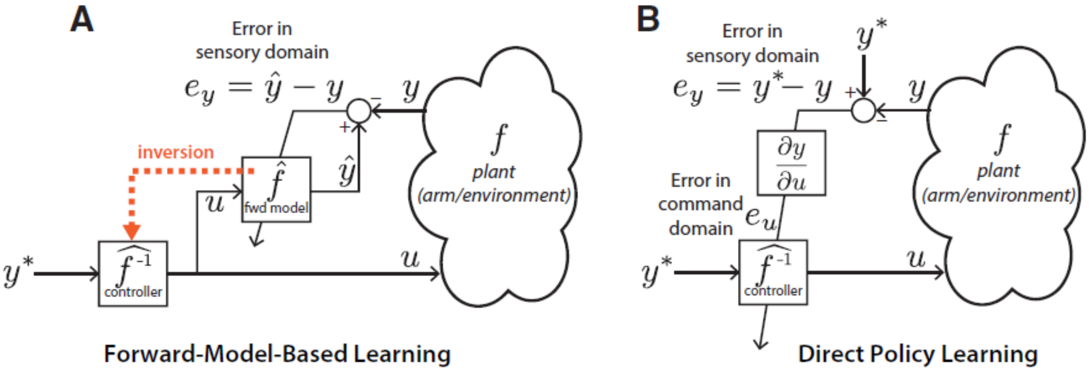
I agree there are important differences between the 'throughput' cases and the decoupled cases. If the literature I cited is on the right track, then an offline model in the medial temporal lobe accounts for some forms of spatial navigation. This makes the system's behavioural capacities considerably more sophisticated. But our question is whether it also makes a difference to the nature of content. My somewhat surprising conclusion is that what makes something a representation in the decoupled cases – what gives content its explanatory bite there – also exists in the simpler throughput cases.

To take an analogy, we need general relativity to explain the perihelion of Mercury but not the trajectory of a cannonball on earth. That does not show that the structures described by general relativity are absent on earth. Representation gets its explanatory purchase in the offline cases, roughly speaking, because representations are calculated over in the service of performing a task function. That also exists in the online / throughput cases, where representations are direct causal intermediaries between input and behaviour, even if, in some of these cases, representation affords little explanatory benefit.

I agree with Krakauer that theorists have often posited unnecessary cognitive sophistication. In a series of important contributions, he has shown how model-free systems can account for subtle forms of motor control. In a recent high-profile paper, he and his collaborators reported impressive experiments which support a model-free account (direct policy learning) over the dominant forward-models paradigm

(Hadjiosif et al., 2021). What is important for my purposes, however, is that forward-model-based learning and direct policy learning both involve computations over internal states (see Figure 1). We need to mark a substantial difference between the two solutions, but it is not the difference between the representational and the non-representational.

< INSERT FIGURE 1 ABOUT HERE >



Forward-Model-Based Learning

Direct Policy Learning

Figure 1. Two rival accounts of motor control, model-based (A) and model-free (B). (From Hadjiosif, Krakauer & Haith, 2021.)

Krakauer I think accepts that model-based systems could be subpersonal in my sense. His objection to my reliance on models in the rat hippocampus is an empirical one – that there is little evidence that the observed patterns of offline replay (during memory consolidation) or preplay (at the time of choice) are the causal basis of navigation behaviour. New evidence in humans does point in that direction (Liu et al., 2020). But I can agree with Krakauer that the evidence is not yet conclusive. Rather than cognitive maps, the behaviour may instead just draw on the 'successor representation' (Momennejad et al., 2017). From my point of view, the clue is in the name. This is not merely a thin use of 'representation' to mean 'decodable' (although that is doubtless one way the term is used (Kriegeskorte & Diedrichsen, 2019)). Successor representations are internal states connected with and computed over in performing a behavioural function. An implicit appeal to function is also apparent in the other examples from neuroscience Krakauer cites (although my account doesn't require that neuroscientists realise that representation is constitutively tied to function).

But is it thinking? We may indeed need a different account for representational content in system 2 thinking. But if so, we will need it for much system 1 thinking as well. Kahneman's examples of system 1 processes involve inputs and outputs that are verbally expressed and conscious (Shea & Frith, 2016). They may also be subject to intra- and inter-personal norms of justification, reasoning and reason-giving, norms which may affect their content. I was never aiming to account for these personal level contents. As to whether my kind of account can be extrapolated to these cases, I am genuinely unsure. My aim was to get clear about at least some kinds of representation, to give us a fixed point around which to build. Cognitive science now abounds with examples where subpersonal representations have the kind of rich model-based structure that Krakauer thinks is needed for RTM. Perhaps surprisingly, the features that make contents so explanatorily efficacious there are also realized in many simpler systems lacking offline use. Both have representations. The functionally more sophisticated systems can do much more with them.

Having representations is not an automatic consequence of performing task functions. It looks is if the poor pithed frog qualifies as performing a robust outcome function. The lecture Krakauer cites suggests that the frog can wipe away irritants at a range of body locations, as well as with different effectors. It is a further, substantial question whether they achieve that outcome in reliance on internal states bearing exploitable relations to distal features of the environment. There are many robust outcome functions in nature, only some of which draw on representations (p. 52). If pithed frogs simply have a collection of direct-throughput mechanisms, then they have achieved robustness in another way. (That is so even if the circuits are hierarchically organised so that when the first one is unavailable a second one kicks in.) One virtue of abandoning the quest for a behavioural criterion for the existence of representations is that there is an empirical distinction between representational and non-representational ways of producing the very same outcome.

Why bother to posit representations where they have little or no explanatory value? The answer is that representations are not posited, they are discovered. We see that a certain kind of pattern arises systematically in nature. We start to understand why. Very roughly, stabilising processes are a diachronic force for producing outcomes robustly, and one way to achieve that synchronically is to calculate over internal states bearing exploitable relations to various features of the problem space. That pattern is found in some quite sophisticated cases involving models and offline calculation. Representational content may be the only practical way to predict and explain behaviour in such cases. But the same pattern, driven by the same process, has arisen in simpler forms. They have representations too, even though there may be no need to appeal to content in explaining their behaviour. Compare: we don't need to appeal to natural selection to explain why there are more dark-coloured moths in a forest of dark tree trunks. We can explain that directly in terms of predation. But natural selection is going on there nevertheless.

Krakauer raises a great issue for philosophers. It presses the liberality objection against my account in a very precise way. It also highlights deep a question, which philosophers are only just beginning to grapple with, about how to distinguish model-based from model-free control (Shea et al., 2008; Butlin, 2021). But it focuses on something that is epistemically useful for attributing representations, rather than metaphysically fundamental to the existence of representations. Speaking broadly, representation is a matter of calculating over internal components in the service of performing a task function. We see, in many case studies, how that can happen online, sometimes in complex ways, as well as in the offline, model-based examples where it is more obvious.

Acknowledgements

I am very grateful to all three commentators for their reviews, for comments on my reply, and for helpful discussions; also to Aleta Quinn for choreographing the process. This research has been supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681422 (MetCogCon).

References

Butlin, P. (2021). Cognitive Models Are Distinguished by Content, Not Format. *Philosophy of Science, 88*(1), 83-102.

Camp, E. (2015). Logical Concepts and Associative Characterizations. In E. Margolis & S. Laurence (Eds.), *Conceptual mind: New directions in the study of concepts.* (pp. 591-621). London / Cambridge MA: MIT Press.

Godfrey-Smith, P. (2017). Senders, receivers, and symbolic artifacts. *Biological Theory, 12*(4), 275-286.

Hadjiosif, A. M., Krakauer, J. W., & Haith, A. M. (2021). Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning. *Journal of Neuroscience, 41*(12), 2747-2761.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual review of neuroscience, 42*, 407-432.

Liu, Y., Mattar, M., Behrens, T., et al. (2020). Experience replay supports non-local learning. *bioRxiv*.

Momennejad, I., Russek, E. M., Cheong, J. H., et al. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour, 1*(9), 680-692.

Shea, N. (2015). Distinguishing Top-Down From Bottom-Up Effects'. In S. Biggs, M. Matthen, & D. Stokes (Eds.), *Perception and Its Modalities* (pp. 73-91). Oxford: OUP.

Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: the case for 'Type Zero'cognition. *Neuroscience of consciousness, 2016*(1).

Shea, N., Krug, K., & Tobler, P. N. (2008). Conceptual representations in goal-directed decision making. *Cognitive, Affective, & Behavioral Neuroscience, 8*(4), 418-428.

Usher, M. (2001). A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation. *Mind & Language, 16*(3), 311-334.

Whittington, J. C., Muller, T. H., Mark, S., et al. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell, 183*(5), 1249-1263. e1223.