Representation in Cognitive Science

A guide for students

Mark Sprevak

25 March 2023

Introduction

Shea wants to defend the *representational theory of mind* (RTM):

- Some of our mental life consists in the processing of subpersonal representations. These representations are *vehicles* – concrete physical particulars inside the brain – which have *content* – states of affairs, typically lying the brain, that the representation is *about*, or that that it *refers* to.

Cognitive neuroscience is full of ideas about:

1. What the relevant vehicles are (what kinds of neural states are representational vehicles),
2. What their contents are (what is represented by neural activities in the brain),
3. Which processes they get involved in during thought (which brain-based computations occur).

However, cognitive neuroscience is less informative about *how the vehicles get their content*. This question is the focus of Shea's book.

1 Existing approaches to the problem of content

There are four approaches in the philosophical literature that are popular when trying to answer this question. All four face serious objections. For a summary of these four approaches, and the problems they face, see the slides from week 1.

Shea claims that the problems can be overcome if one combines elements from each of these four approaches in the right way. The appropriate blend of elements might vary across different cases in cognitive neuroscience (this is what he calls 'pluralism'). In the rest of the book, he will sketch how this all works.

2 The ascriptionist approach to the problem of content

Shea mentions a fifth approach to the problem of content: the *ascriptionist* approach. He places Dennett and Davidson under this heading.

According to an ascriptionist approach, a mental representation is ascribed to a whole person based on how they behave. For example, on Dennett's view – the intentional stance – the content of your personal-level *beliefs* and *desires* is determined by what would best explain the patterns in your behaviour. If we adopt the intentional stance towards you, we assume you are a rational being and attribute to you the beliefs and desires that would best explain and predict your behaviour.

A key characteristic of this view is that it does *not* commit to these mental representations being concrete physical particulars inside your head. In fact, Dennett pours scorn on that assumption – calling it 'industrial strength realism' about mental representations. If you want to know more about his reasoning here, read this classic article (posted in the 2ndary readings):

- Dennett, D. C. (1978) 'A Cure for the Common Code' in *Brainstorms*, Montgomery, VT: Bradford Books, pp. 90–108

Shea claims that because ascriptionist approaches do not support the idea that mental representations are concrete particulars inside the brain, they are *not* a suitable way of defending RTM.

Hence, he will not consider ascriptionist approaches further.

NB. Dennett believes in the RTM – he thinks that the intentional stance only applies to *personal-level* representations. It does not apply to subpersonal representations. On his view, some other (non-ascriptionist) story most be told about how subpersonal representations get their content. Hence, Dennett and Shea could, in theory, agree on what follows in the book.

Framework

1 Setting aside harder cases

Shea wants to focus on *subpersonal* mental representations. He does not define precisely what subpersonal representations are (although he will give many examples in the book). He does not propose any general theory of the personal/subpersonal distinction (see the uploaded readings for some suggestions here). (NB. Shea does not endorse the popular view that personal=whole organism and subpersonal=proper parts of organism).

Shea uses the term 'subpersonal representation' to signal that he will set aside certain kinds of representation for which the determination of their content is likely to depend on factors other than those he considers in the book:

1. Consciousness
2. Justification and satisfying certain evidential epistemic relations
3. Its role in reasoning-giving explanations
4. Having certain natural-language-like structural features

So what then are subpersonal representations?

For Shea, subpersonal representations are *neural states* that play a distinctive computational role in generating our behaviour and cognitive processes and moreover that they have some *determinate representational content* that plays a role in explaining the success or failure of that behaviour.

2 What should constrain our theorizing?

How do we test our theory of representational content? How would we know if we have got it right?

Shea argues that our pre-theoretic intuitions shouldn't be given any weight.

Rather, we should look at the *practice of cognitive science*. Specifically, we should look at cases in which cognitive science attempts to explain an organism's *behavioural successes and failures* in terms of its subpersonal representations.

If we, as philosophers, get the facts about a theory of content right, then we can make sense of how attribution of *those* subpersonal representational contents provides a good explanation of its behavioural successes and failures – a *better* explanation than one that attributed *different* representational contents or one that *did not attribute any representational contents at all* (a 'factorised' explanation). What getting the theory of content right means is to vindicate the success of such representation-involving explanations in cognitive science:

> An account of how representational content is constituted in a class of systems should allow us to show why recognizing the representational properties of such systems enables better explanations of behaviour than would be available otherwise. (p. 29)

3 Task functions and algorithms

Shea gives some examples of *tasks* – or what he calls *task functions* in this chapter.[1] Tasks describe the various problems that an organism encounters as it goes about its business. They break down the challenges the organism faces into a series of discrete issues. Relative to those tasks, one can make sense of specific instances of the organism's behaviour counting as examples of *success* or *failure*. Success = achieving the mapping described by the task; failure = not achieving this mapping.

A task in this sense should be understood in black-box, purely input–output terms:

- For a maze-solving task: the task might be a mapping from the organism's current location to its desired location
- For a coat-making task: the task might be a mapping from raw materials to a completed coat
- For an arithmetic task: the task might be a mapping from a long written multiplication problem to the correct written answer

How does an organism try to solve its tasks?

Often the answer is via an *algorithm*. Shea characterises an algorithm as a step-by-step mechanical procedure involving representations. He glosses this as a *computation* over representations. A computational process is only sensitive to the intrinsic properties of the representational vehicles it manipulates.

4 Vehicles and syntactic types

Shea finesses his view: the computation is not best characterised over vehicles but over *syntactic types*.

Shea defines *vehicles* as the concrete particulars that bear representational content. Two similar shaped sets of ink marks – 'barn' and 'barn' – count as two vehicles of the same vehicle type. In general, vehicles are individuated by their intrinsic physical/functional properties. This means that if two vehicles share their same intrinsic physical/functional properties, then they are the of the same vehicle type.

However, there is a problem. Two vehicles that are exact duplicates in terms of their intrinsic physical/functional properties might have *different* representational content. For an English speaker, 'barn' represents *barn*; for a Swedish speaker, the same same vehicle 'barn' represents *child*.

To get around this difficulty, Shea introduces the notion of a *syntactic type*. A syntactic type is a way of grouping vehicles together that allows us to say that they share the same representational content.

How then should we group vehicles together into syntactic types? Shea proposes that syntactic types should be individuated by *how vehicles are processed*. If two vehicles are processed in the same way by the organism, then they count as being of the same syntactic type. Algorithms are therefore defined over representational vehicles of the same syntactic type.

[1] Slightly confusingly, Shea appears to define the term 'task function' differently in the next chapter. There, he says that it refers not to a mapping between input and output, but to a specific type of behavioural output – a stabilized, robust behavioural output. The task function of the heart, for example, is to *pump blood*. The task function of a chickadee when scavenging is to *reliably retrieve cached food*.

(NB. Shea can't define syntactic types in terms of the vehicles having the same representational content, because he wants to use syntactic types to explain content, not vice versa!)

5 Pluralism and varitel semantics

Shea's aim is to explain how facts about semantics (representational) facts arise from non-semantic (physical/functional) facts. He suggests that there are *multiple* ways in which this could happen (at least 8!). His position is 'pluralist'. Shea admits that there is a family resemblance between the 8 ways he describes, but he is not committed to them all being instances of one single physical/functional condition.

Shea calls his position 'varitel' semantics. The name suggest is that there are *various* ways in which a teleosemantic condition on representation can be fulfilled.

How do the 8 ways he considers arise?

- There are 4 ways in which a task an organism faces might acquire a metric of *success* or *failure*:

    1. evolution by natural selection;
    2. learning from feedback;
    3. contribution to the organism's survival/homeostasis;
    4. deliberate design.

- There are 2 relationships that a state inside the organism might bear to its environment that might be exploited by the organism to achieve behavioural success on tasks:

    1. carrying correlational information;
    2. structural correspondence.

Functions for representation

1 Overview of chapter

Shea's account says that a subpersonal representation is a concrete internal state that *explains patterns of success and failure in our behaviour*.

This raises two questions:

1. How does a metric of success or failure come to be attached to instances of our behaviour?
2. How do subpersonal representations explain those patterns of success and failure?

Shea needs *naturalistic* answers to both these questions.

This chapter concerns how he answers (1). The following 2 chapters look at how to answer (2).

An organism's physical behaviour, in and of itself, has no objective measure of success or failure attached. No behaviour is 'successful' in and of itself. We need to look to somewhere beyond the physical mechanics of behaviour to find the facts that determine that a particular instance of behaviour counts as successful or not. Where do the *norms* of success/failure arise from? And do they depend on purely naturalistic facts?

(NB. This is a long-standing issue in philosophy of biology. It is normally phrased in terms of the problem of how to account for *biological functions* in naturalistic terms. How do we characterise what a behaviour or biological trait is *for*? There are lots of attempts to try to answer this question by appealing to the past or future reproductive/survival success of that behaviour. For an introduction, see this REP entry by Karen Neander).

Shea provides a disjunctive answer to the question where the relevant 'success' standards come from for behaviour. He suggests that *a standard for behavioural success* arises from 3 possible sources. All of these concern the *past history* of the organism:

1. Behaviour of this type in the evolutionary past contributed to the organism's ancestors' reproductive success/fitness
2. Behaviour of this type in the organism's own past was reinforced by learning
3. Behaviour of this type in the organism's own past contributed to the persistence of the organism

'Successful' behaviour in this context means behaviour that fulfils one or more of conditions (1)–(3). Subpersonal representations explain how the organism often produces patterns of success and failure.

(1)–(3) sometimes all agree about which behavioural outcomes count as 'success'. However, sometimes they don't. A behaviour might count as 'success' relative to one standard, but not relative to another. For example, a male spider's mating behaviour, which results in it getting eaten by a female, might count as a 'success' relative to standard (1) but 'failure' relative to standard (3). On Shea's account, we need to be aware that there are *different* things that a cognitive scientist might mean when they talk about 'success'.

Shea focuses on (1)–(3) because they are, in his view, the most likely candidates for providing naturalistic standards for behavioural success/failure.

He briefly touches on a 4th standard for behavioural success/failure. This is *intentional design*. We might, as either designers or users, *intend* that certain behaviour of an organism counts as 'successful' and other behaviour does not. Our intentions can certainly legislate a standard for success, but it does not introduce a naturalistic standard. For what counts as success/failure according to this metric depends on facts about our intentions, beliefs, values, interests, etc.

2 Task functions, robust functions, stabilized functions

Let's dive into how Shea talks about the issues. Shea introduces three terms in the chapter:

- Robust function
- Stabilized function
- Task function

We will look at these more closely in turn.

In all cases, please be aware that the word 'function' refers to a kind of successful *behavioural outcome* for the organism. The function of the heart, in this sense, for example, is to *pump blood*. The function of a chickadee when scavenging is to *reliably retrieve cached food*. The function of the toy system described on p. 67 is to *reach point T*.[1]

When Shea says that the aim of the chapter is to provide a naturalistic account of *functions*, what he means is the aim of the chapter is to provide a naturalistic account of *what makes for a successful behavioural outcome* for the organism.

On Shea's view, a behavioural outcome might be a physical movement (e.g. moving your eyes 12 degrees to the right), or a consequence of the organism's actions (e.g. winning £50, getting a pool ball into a pocket). He is not concerned with individual tokens of behavioural outcomes but rather with the production of successful types of behavioural outcome.

2.1 What is a robust function?

A *robust function* is a type of behavioural outcome that is produced by the organism across a wide range of different sensory inputs, and across a wide range of different intervening external circumstances.

An example of a robust behavioural outcome for a squirrel might be *getting a nut from a bird feeder*. The squirrel will tend to get a nut from the feeder under a range of different sensory inputs – it will get a nut whether it approaches the feeder from one direction or another, whether it sees it on a cloudy or sunny day. The squirrel will also get the nut under a range of intervening external circumstances – it will tend to get it regardless of how strong the wind is blowing or whether you have placed an 'squirrel-proof' collar around the feeder.

What sorts of changes in conditions should be considered here? No behavioural outcome is robust under *all* circumstances (e.g. an asteroid hitting the squirrel will certainly stop it getting the nut). Similarly, if the squirrel were to only get the nut under variations of ±0.00001C in temperature that would not thereby make the outcome robust. Shea says that we should look for robustness under 'relevant' changes to the conditions. But what counts as 'relevant'? This needs to be cashed out in a fully naturalistic way ('relevant' can't mean interesting to us!). Shea outlines a few ideas about how do this on p. 56, but it's not clear the issue is completely addressed.

[1] This is not what is meant by 'function' in mathematics or computer science, where a function is a mapping or pairing of two things. There is a potentially confusing double usage of the term in Shea's earlier chapter.

A robust outcome behavioural outcome may not always map onto a 'successful' behavioural outcome. Unsuccessful outcomes can also be robust – think of someone's robust ability get a low mark on an advanced maths assignment (no matter how they view the problems or the intervening circumstances). To identify which robust behavioural outcomes are more/less successful, Shea introduces the notion of a *stabilized* function.

2.2 What is a stabilized function?

A stabilized function is a type of behavioural outcome that has been 'stabilized' by one or more of the following mechanisms:

1. Behaviour of this type in the evolutionary past contributed to the organism's ancestors' reproductive success/fitness
2. Behaviour of this type was reinforced by learning (during the lifetime of the organism)
3. Behaviour of this type contributed to the persistence of the organism

What does it mean for a type of behavioural outcome to be 'stabilized' by one of these mechanisms?

It means that an organism producing that behavioural outcome *increases the chances* of the organism (or its descendants) producing a behavioural outcome of that same type in the future. Stabilized behavioural outcomes *entrench* themselves – their occurrence makes a future occurrence of the same type of behavioural outcome more likely.

The three mechanisms (1)–(3) are distinct (but interrelated) ways in which this kind of entrenchment/stabilization of behaviour can occur:

- For (1), if behaviour of that type contributed to an organism's ancestors' fitness in the evolutionary past, then it more likely – due to the way that natural selection works – that behaviour of that type will be produced by the organism in the future.

- For (2), if behaviour of that type was rewarded (reinforced) by a learning process in the past, then it is more likely – due to the way in which reinforcement learning works – that behaviour of that type will be produced by the organism in the future.

- For (3), if behaviour of that type contributed to the persistence of the organism, then it is more likely – as the organism is around to keep doing it! – that behaviour of that type will be produced by the organism in the future.

The three mechanisms (1)–(3) need not be present in all cases, but often two or more of them act in unison to stabilize a behavioural outcome. For example, a monkey plucking ripe red fruit might be a type of behaviour that was: (i) selected for during the monkey's evolutionary history; (ii) rewarded during the monkey's own lifetime; and (iii) in the past contributed to the persistence of the monkey.

Note that Shea's account the facts about stabilization are entirely *backwards looking*. He defines stabilization in terms of *what actually happened in the past* (the organism's own past as well as its evolutionary history). Shea contrasts this with 'forward looking' accounts of biological function. Observe that his characterisation of stabilized functions means that, for Shea, 'swamp systems' will, at their moment of creation, have no stabilized functions (and hence no representational content), because they have no history.

What is the relationship between stabilized outcomes and robust outcomes?

Shea argues that stabilization is one way in which a system may come to produce robust behavioural outcomes. It is not the only way, however. Another possible source of robustness is *intentional* *design*: *viz.* us engineering the system to behave in certain ways under a wide range of conditions. Shea claims that the stabilization mechanisms (1)–(3) are 3 possible *naturalistic* sources of robust behavioural outcomes.

2.3 What is a task function?

A *task function* = a robust, stabilized function.

(Shea also says that robust functions produced *via intentional design* count as task functions (p. 65). However, since they are not naturalistic task functions, so we won't be concerned with them in the rest of the book.)

In the context of the book, the point of task functions – robust, stabilized behavioural outcomes – is that they pick out the 'successful' behavioural patterns that Shea wishes to focus on. Note that Shea is taking great pains to pick out this subset of behavioural outcomes purely naturalistically: he is focusing on *robust*, *stabilized* outcomes (as those two terms are defined above). He is not assuming any additional external normative standard of 'success'.

Shea will claim that explaining how an organism accomplishes its task functions synchronically – explaining how, in any given instance in the here and now, it produces this (robust, stabilized) behavioural outcome – frequently relies on a causal story about computational relationships between concrete states inside the organism that bear certain special ('exploitable') relations to environmental states. According to Shea, *that is all there is to subpersonal representations.*

If an organism accomplishes its task function in this way, those internal states – which causally drive its behaviour – *are* its subpersonal representations. The content of those subpersonal representations is determined by looking at (i) the corresponding task function and (ii) the corresponding exploitable relations.

3 The Robust, Stabilized, IntC cluster

RBST Robust behavioural outcomes
STAB Stabilized behavioural outcomes
INT.C. Behavioural outcomes produced as a result of interaction of internal components bearing exploitable relations to the environment

Shea claims that these three properties tend to be found together in nature.

Often, a robust behavioural function will have been produced by a stabilization mechanism (1)-(3). Robust behavioural outcomes (Rbst) thus tend to often co-occur with stabilized behavioural outcomes (Stab). Note, however, that the two do not *always* or *necessarily* co-occur.

Stabilization concerns the question of how – over evolutionary time or the lifespan of the organism – a robust behavioural outcome has become entrenched. There is also the question of how, on any given occasion, this robust, stabilized behaviour is generated by the organism. This concerns what Shea calls the 'synchronic' mechanism that generates behaviour in the here and now – as opposed to the 'diachronic' mechanism working over longer periods of time that entrenches certain behavioural outcomes.

Shea claims that, *in many cases*, the synchronic mechanism for producing a robust behavioural outcome consists in an interaction of internal states that stand in exploitable relations to the environment (Int.C.). An internal process (an algorithm) manipulates synchronically structured internal states in the organism that stand in special 'exploitable' relations to distal states in the environment.

Shea notes that Rbst, Stab and Int.C. can and do come apart in the real world – they do not always or necessarily co-occur. However, he claims, it is *often* the case that they come together in natural systems. When they do, Shea's conditions for subpersonal representations are satisfied.

Correlational information

In the next 2 chapters, Shea looks at 2 forms of 'exploitable' relationship between internal states in the organism and external features in the environment. Either one or both might be an ingredient in a case of representation. The two relationships he considers are:

1. Exploitable correlational information
2. Exploitable structural correspondence

In this chapter, he focuses on correlational information.

1 Correlational information

What is correlational information?

Shea defines it in terms of probabilities. Let us call:

- Fa: Some internal vehicle a being in state F
- Gb: Some external environmental state b being in state G

Then internal vehicle a being F carries *correlational information* about external environmental state b being G just in case:

- $P(Gb \mid Fa) \neq P(Gb)$.

In other words, Fa carries correlational information about Gb provided events Fa and Gb are not probabilistically independent of each other.

This is the kernel of Shea's idea about Fa representing Gb. Intuitively, one might imagine that an organism might condition its behaviour on Fa as a way to condition its behaviour on some Gb. It might use (easily accessible) state Fa as a proxy for – as a way of interacting with – hard-to-reach, distal environmental state Gb.

However, there is a big problem with simply applying this unvarnished kernel as an account of representation. It is an extremely weak condition. Few events in the real world are probabilistically independent of each other. Equating representation with correlation would yield a wildly liberal account of representation; one so liberal as to be useless to cognitive science. In the rest of the chapter, Shea tightens up the condition by layering on further requirements.

2 *Exploitable* correlational information

First, Shea lays on three further requirements that characterise a notion that he calls *exploitable* correlational information. These are:

1. The correlation between Fa and Gb should not be a 'one-off', a singular occurrence. It should hold for across a range of different conditions (perhaps over a region space, over a period time, over various different choices of a, b, etc.). This is fleshed out by Shea's talk of the correlation holding across 'regions', D and D'.

2. The correlation should hold *non-accidentally* – it should somehow be counterfactually stable and governed by laws of nature (nomologically supported). This suggests that there should

be a modal dimension to the correlation – it should cover, not only what actually happened, but also what would have happened if things had been different.

3. The correlation should hold for some *univocal* reason. There should be a single reason why it holds across the regions, e.g. some factor in common across different cases that explains it holding. (Shea gives the example of green-123 being poisonous in meat and veg for different reasons as a correlation that would violate this condition).

If (1)–(3) are met, then we have what Shea calls *exploitable* correlational information.

But exploitable correlational information is still an extremely liberal notion. Exploitable correlational information as defined above abounds in the world. The definition places no limits on the size or selection of regions D and D' (which could be as small or gerrymandered as you like). It also places no limits on the strength of correlation required (it just needs to above zero). Shea spends the next sections adding further requirements to make it a more suitable stand-in for the notion of representation as used in cognitive science.

This is where we hit the meat of the view.

3 *Exploited* correlational information ('UE information')

According to Shea, in order for a state to count as a representation, it is not enough for that state to *have* exploitable correlational information; the correlational information has to actually *be exploited* by the organism in some appropriate sense.

A lot of ideas are packed by Shea into this latter condition, which he labels carrying 'UE information'.

We are going to look at the condition in detail. However, to give you brief preview of where we'll get to, what Shea intends by a correlation being exploited by the organism is that:

- The correlation played some sort of causal-explanatory role in the organism's success (namely, in achieving stablized, robust outcomes). The correlation should be of an appropriate kind to *causally explain* the robust, stabilized behavioural outcomes produced by the organism.

Let's back up a little and unpack Shea's proposal step by step.

3.1 Constraining regions, constraining correlation strength

First, as noted above, the bare notion of exploitable correlational information places no constraints on the choice of regions D and D' or the strength of the correlation.

Shea suggests that for a correlation relevant to representation:

1. The regions, D and D', should include those regions in which the stabilization and robust outcomes behavioural outputs for the task function occur.
2. The correlation should be strong enough to causally explain the occurrence of the stabilization and the robust outcomes.

3.2 Causal explanation

A crucial ingredient in Shea's account of representation is *causal explanation*. Certain correlations count as representations because they *causally explain* behavioural successes (stabilization and robust outcomes).

It is very important to be aware of two points here.

3.2.1 *Two types of explanation*

Two types of explanation are discussed by Shea in the book: *representational* explanations and *causal* explanations. Representational explanations explain behaviour of an organism in terms of processes involving representations inside the organism. Causal explanations explain its behaviour in terms of causal relations between physical states.

The two forms of explanation explain by appeal to entirely *different* kinds of explanans: representational explanations appeal to representational states; causal explanations appeal to causal relations and physical states.

Shea is *not* proposing that the facts about whether a state is a representation depends on that state's role in a representational explanation. That is absolutely not his project at all. If he were to do this, it would unclear how he would be offering an account of representation in terms that did not already presuppose representation.

Shea is proposing that whether a state counts as a representation depends on that state's role in a *causal* explanation of behaviour.

3.2.2 *Causal explanations are objective facts*

You might worry that what counts as a 'good' causal explanation is a subjective matter. Whether an internal state features in a causal explanation of that organism's behaviour depends on our interests and attitudes – on what *we* find satisfying to learn about. Isn't what counts as a good explanation just a matter of what fits with our psychological needs?

It is very important for Shea that the answer to this question is 'no'. On his view, it is an *objective matter* whether a causal explanation of behaviour is a good or bad one. Whether A explains B is not a matter of what we find psychologically satisfying, that quenches our understanding, etc. It is a matter of certain objective facts obtaining. This enables Shea to naturalise the facts about representation in terms the facts about causal explanation.

In making this assumption, Shea follows a long tradition in philosophy of science that treats explanation as an objective relation between an explanans and an explanadum. Shea does not discuss any specific proposal here, but there are plenty (starting with Carl Hempel). Shea's claim is that *if* there are objective facts about what counts as a good causal explanation of behaviour, *then* there are objective facts about the representations the organism has.

(To learn more about objective accounts of scientific explanation, see the review paper by Brad Skow on the reading list, and the Stanford entry that reviews so-called 'pragmatist' approaches to explanation, which argue that explanation is *not* an objective relation.)

3.3 **Putting the pieces together: UE information**

We are now in a position to put the pieces together.

The explanadum that Shea identifies consists of two elements:

1. How was the behavioural outcome stabilized (over the past)?
2. How are robust behavioural outcomes produced (now)?

Shea places both under the heading 'explanation of the task function'.

Let us return to our question: What does it means for correlational information to be exploited by the organism (rather than just being exploit*able*)?

(E) Correlational information is exploited by the organism just in case it plays an unmediated role in a causal explanation of the task function.

The only part left to be spelled out is the 'unmediated role' condition. This means that the role of correlation in explaining the behaviour does not rely on appealing to a further correlation. (Ruling out explaining success in catching flies by appeal to correlation between a neural state and black dots on the retina plus a further correlation between those black dots on the retina and the presence of flies).

If Shea's condition (E) is met, then he says that the internal state in question *carries unmediated explanatory information* (UE information) about the distal environmental condition.

The expression 'UE information' is rather a mouthful. The important thing to remember is that UE information involves a fairly hefty condition involving causal explanation being met. *That* is how Shea gets around the problem of representation not being mere correlation.

4 Representation via correlation information

The upshot from the chapter is a sufficient condition for a component state *R* of an organism to be a representation with content *C*:

- *R* carries UE information about *C*

Parsed into more friendly language:

- A sufficient condition for *R* representing *C* is that there is a correlation between *R* and *C*, and that correlation features in the causal explanation of some of the organism's behavioural successes (understood in terms of stabilized, robust behavioural outcomes).

That's it!

In the remainder of the chapter, Shea argues that his proposal fits with a number of case studies in cognitive science and explores some of its consequences.

One consequence to highlight is that Shea says his 'good causal explanation' condition will rule out weird, disjunctive states featuring as representational contents (p. 90). Weird, gerrymandered distal states (e.g. *cow-or-horse-on-dark-night*) can have very high degrees of correlation with internal states. But, Shea claims, since they are 'nonnatural' states, they are not good candidates to feature in causal explanations, and hence not good candidates for representational contents. This is how Shea tries to solve the disjunction problem (see week 1).

Structural correspondence

In this chapter, Shea looks at a second form of 'exploitable' relationship between internal states in the organism and external features in the environment. He focuses on exploitable *structural correspondence*. According to Shea, the right kind of structural correspondence between internal states in the organism and environmental states simply *is* representation – just as, in the previous chapter, he argued that the right kind of correlation between internal states in the organism and environmental states simply *is* representation.

Note that Chapters 4 and 5 concern different kinds of representation:

- In Chapter 4, the question was whether a *discrete* internal vehicle represents some distal environmental content. In theory, that internal vehicle does not need to have any internal structure. It might be an atomic state for the organism.
- In Chapter 5, the question is whether a *structured* collection of internal states represents some collection of distal states and relations in the environment. Such a collection cannot be an atomic state – it must be made up of smaller parts and relations.

You can think of this as like the contrast between representations like *words* – which carry information about the environment without themselves needing to have an interesting internal structure – and *sentences* – which also have an internal structure with multiple constituent vehicles (words) and relations (ordering of the words). This structure sentence often represents more than what is conveyed the individual words. For example, 'Mary stands behind John' represents a different environmental state to 'John stands behind Mary', even though they both use the same words.

Note that there is nothing to prevent an organism using *both* UE information *and* structural correspondence to achieve its behavioural success. The constituent parts of a structured representation often also carry UE information about the environmental states to which they correspond (see rat place cells, pp. 114–115). But that condition may not always be met (see the example of icons on p. 117). On the general relationship between UE information and structural correspondence – it's complicated. Shea does not attempt to fully map this out. The main thing to remember is that it is not a case of an either/or choice here.

For the rest of this chapter, we will focus just on structural correspondence.

1 Structural correspondence

As in Chapter 4, Shea starts with a very thin notion that would result in an implausibly liberal naturalistic theory of representation and gradually builds in extra conditions until it becomes a more plausible candidate naturalistic condition for representation. In the previous chapter, he started from the thin notion of *correlation*; in this chapter, he starts from the thin notion of *structural correspondence*.

What is this thin notion of structural correspondence?

Intuitively, you can think of it as a kind of 'mirroring': where the states and arrangement of states inside one domain mirror the states and arrangement of states inside another domain. In this chapter, the two relevant domains are *what goes on inside the head* and *what goes on outside the head*. The question is whether an organism's neural states and relations systematically map to external environmental states and relations.

Let's make this more precise. The two domains are:

1. Internal states (v_i) and relations between them (V) inside the organism
2. External states (x_i) and relations between them (H) in the distal environment

Note that f is just some purely formal mapping relation between elements of the two domains. It is nothing to do with correlation or carrying UE information.

A structural correspondence is defined as a *structure-preserving mapping from domain (1) to domain (2)*. A structure-preserving mapping is a 'translating' function f – some scheme for mapping, or 'key' – that would take one *from* the states inside the head (v_i) *to* the states outside the head (x_i) such that relationships between the corresponding states are always preserved.

More precisely, a structural correspondence holds between the domains iff:

(C) V holds between $v_i, v_j \Leftrightarrow H$ holds between their external counterparts, $H(f(v_i), f(v_j))$

You might want to pause here and apply this formalism to some examples to convince yourself that it captures the intuitive notion of mirroring.

Can we simply stop here and say that if (C) is true, then we have a case of structural representation?

Unfortunately, no. (C) is *trivial* to satisfy. Provided there are enough internal states, v_i – and provided no constraints are placed on our choice of relation V – then a structure-preserving mapping, f, will *always* exist. A structural correspondence exists between any two domains with enough elements.

(If that result doesn't seem obvious, you are in good company. Bertand Russell fell into this trap with his structural-based theory of knowledge. Max Newman published an elegant criticism in *Mind* in 1928 that showed Russell's condition for knowledge was trivial. Have a look at pp. 112–113 of the book where Shea shows the relevant reasoning with a simple example.)

The root cause of the problem is that (C) places no constraints on the choice of V. That leaves the door open to choosing a bizarre V that the organism makes no use of but that structurally corresponds to any arbitrary H one likes. (Equally, there is no restriction placed on H, so one is free to pick any V inside the organism and then find a weird, gruesome H in the environment to which it corresponds – and voilà, a structural representation!)

In the rest of the chapter, Shea adds further conditions to (C) that restrict V and H in fairly hefty ways, and thereby provides a non-trivial condition of structural correspondence. Shea argues that if this more substantial form of structural correspondence is met then, then one has structural representation.

2 *Exploitable* structural correspondence

As a first step, Shea adds two extra conditions that characterise what he calls *exploitable* structural correspondence:

1. V is a relation to which processing in the organism is *sensitive*
2. H and x_i are *of significance* to the organism

Let's take these one at a time.

Condition (1) says that processing in the organism should be sensitive to relation V. What does that mean?

Shea unpacks it by saying that V should make a systematic difference to downstream processing in the organism. Intuitively, if V holds between two internal states v_1, v_2, then the organism should be able to detect that and respond differently to a situation in which V did not hold between those two internal states. This sensitivity to relation V should also be systematic – it shouldn't be a one-off or only apply for certain pairs of v_i, v_j and not others. Ideally, processing in in the organism should be sensitive to whether V hold or not across any pair of elements v_i, v_j, and across a wide variety of background conditions. There is plenty more to say here (e.g. how much sensitivity is enough?) but Shea moves on. An example of his intended idea: a rat's downstream processing *is* sensitive to whether the rat's place cells *fire at the same time* but *not* sensitive to whether those cells *have the same colour*.

Condition (2) turns to a restriction on H. It says that the states and relations H and x_i must be of significance to the organism. What does that mean?

Shea unpacks 'significance' in terms of the environmental states and relations being significant to the organism relative to its task function (i.e. to producing its robust, stabilized behavioural outcomes). Slightly strangely, Shea does not say much more, and he does not explicitly define how we should understand significance relative to a task function. What does it mean for environmental states, x_i, and relations, H, to be 'significant' to a task function? Taking inspiration from the previous chapter, and from what comes next in this chapter, perhaps 'significance' should be understood as meaning *featuring in the correct causal explanation for how the organism accomplishes the task function*. That is my guess of what Shea has in mind here.

Conditions (1) and (2) move us out of the worrisome realm of Newman's objection and the triviality result. We are simply no longer free to pick *any* weird V or H we choose.

3 *Exploited* structural correspondence ('UE structural correspondence')

Shea says that the notion of exploitable structural correspondence is still too liberal to underwrite a notion of structural representation. According to Shea, in order for a state to count as a structural representation it needs not only to be *exploitable* (as defined by 1 and 2 above), but also actually *exploited* to accomplish a task function.

Shea's case for the 'exploitable' relation being too liberal is perhaps not so clear as it was in the previous chapter (a useful exercise for you would be to compare the different ways of defining 'exploitable' across the two chapters). Shea makes his case in Section 5.5 where he gives some putative examples of *unexploited* structural correspondence that don't count as representations. For example, he considers the relations between bee dances, which *could* be exploited by bees to find nectar, but isn't exploited by them – bees don't exploit relations between dances (although presumably they can see them), they only respond to the waggles of an individual dance.

For the sake of argument, let's assume that Shea is right that a structural correspondence must be exploited by the organism in order for it to count as a structural representation.

What does Shea mean by a structural correspondence being *exploited*?

Similar to the previous chapter, the key idea is its role in *causal explanation* of behaviour:

(S) An exploitable structural correspondence is exploited by an organism just in case that structural correspondence plays an unmediated role in a causal explanation of the task function.

In plainer English, an exploitable structural correspondence is *exploited* just in case it features in the correct causal explanation of the organism's behavioural success (i.e. it achieving stablized, robust behavioural outcomes).

As in the previous chapter, it is vital that the norms of causal explanation – what counts as the 'right' causal explanation of behaviour – is a purely objective matter. Shea is proposing that the facts about representation depend on the facts about causal explanation, so the facts about causal explanation had better be a 100% naturalistic, objective matter.

The only part left to be spelled out is the 'unmediated role' condition. This just means means that the role of structural correspondence in explaining the behaviour does not rely on appealing to some further structural correspondence obtaining.

Note that instead of the term 'exploited' structural correspondence, Shea prefers to use the term 'unmediated explanatory structural correspondence' (UE structural correspondence). These mean the same thing.

4 Representation via structural correspondence

The upshot from the chapter is a naturalistic sufficient condition for a collection of internal states (v_i) and relations (V) inside an organism to be a structural representation of a set of environmental states (x_i) and relations (H):

- A UE structural correspondence obtains between V and H

Parsed into more friendly language:

- Processing in the organism is sensitive to V; the x_i and H are of significance to the organism; and a structural correspondence between the V and H plays a role in the causal explanation of the organism's behavioural success ('success' understood in terms of the organism producing stabilized, robust behavioural outcomes).

That's it!

Shea explores a range of other quirks and interesting features of the view in the rest of the chapter. These include the question of how an organism might start from a set of internal states that don't quite stand in an exploitable structural correspondence to the environment – they stand in what he calls a *potential* exploitable structural correspondence – and how that might get tuned by learning to become an exploitable structural correspondence over time. He also considers cases of *approximate instantiation* of a structural correspondence: cases in which differences in v_i and V don't exactly track differences in x_i and H, but are still good enough to play a role in causing behavioural success. There is a lot to explore in the chapter, and Shea leaves many issues open for further work.

Standard objections

This chapter looks at objections that have been raised to past philosophical attempts to naturalise representational content. Shea argues that his account is not vulnerable to those objections.

Two main objections considered in this chapter.

1 Objections from indeterminate content

Past attempts to naturalise representational content are often criticised for yielding contents that are implausibly indeterminate. This is seen as a *reductio ad absurdum* of these theories of representation.

Consider the classic frog case. Let R be activity in neurons in the frog's retinal ganglion. Let F be the presence of a fly at a distal location (x, y, z). It is tempting for a simple theory to say that neural activity R *represents* F because R correlates with the occurrence of C.

However, if that simple story is right, then R also represents more than the presence of the fly:

1. **Distality problem**: R also correlates with more proximal events too (e.g. a pattern light falling on the frog's retina). Why does R not represent those as well?
2. **Specificity (aka *qua*) problem**: Ignore the distality problem; R will also correlate with lots of distal things that are present at (x, y, z). In addition to a *fly* being there, there is also a *little black thing*, a *flying nutritious object*, *something worth eating*, *something good for the frog*, … Why does R not represent all those distal states too?
3. **Disjunction problem**: If R correlates with *fly at* (x, y, z), then it will correlate (and even more strongly!) with a disjunctive condition such as *fly or BB pellet present at* (x, y, z). Why does R not represent those disjunctive conditions?

Shea has a two-pronged approach to answering these worries about content.

First step: He points out that his naturalistic theory of content is based on much more than brute correlation. There are numerous extra conditions in his account that need to be met.

Three conditions in particular work to address these indeterminacy problems:

1. (**Distality problem**): For Shea, a UE correlation must play an *unmediated* role in explaining the organism's behavioural success. The pattern of light falling on the frog's retina only explains the frog's behavioural success (*viz.* eating the nutritious fly) if those retinal patterns are themselves are correlated with presence of a distal flies at (x, y, z). The retinal patterns do not play an unmediated role in explaining the frog's behavioural success. Hence, they aren't candidates for representational content.
2. (**Specificity (aka *qua*) problem**) For Shea, a UE correlation must feature in the casual explanation of an organism's behavioural success. In other words, it must contribute to a robust behavioural outcome that that has been stabilized over past history. In the past, eating *flies* contributed to the frog's survival and its ancestors' reproductive success. In contrast, eating *little black things* did not. Hence, the former, but not the latter, is a candidate for representational content.

3. (**Disjunction problem**): For Shea, a UE correlation between *R* and some distal state must feature in the casual explanation of the frog's behavioural success (*viz.* eating the nutritious fly). Disjunctive distal properties – like *fly or BB pellet present at* (x, y, z) – don't normally qualify as candidates for featuring in any good causal explanations. Hence, they are not candidates for representational content.

Second step: Shea acknowledges that some residual degree of indeterminacy in an organism's subpersonal representational content is likely to remain.

On his account, it will be indeterminate whether, for example, the frog's internal state *R* represents *fly* (biological category) or *flying nutritious object* (ecological category). Both distal states stand in appropriate correlations with *R*, both correlations explain the frog's behavioural successes, both seem to be plausible candidates for featuring in a causal explanation (they do not involve disjunctive, gerrymandered, or non-natural states).

Shea argues that this kind of indeterminacy is to be expected and should not be regarded as a problem for his theory. It is just a reflection of the frog not being able to draw a distinction between certain distal states. The case only appears problematic to us because we, as sophisticated language users, notice a more fine-grained distinction between the distal states.

2 Objections from Swampman cases

An important feature of Shea's account is that it makes representational content depend not only on the facts about the organism in the here-and-now, but also on temporally distant facts about that organism's history. This is because, according to Shea, the representations that an organism has now depend on that organism's task functions, and those task functions depend on *which behaviours in the past contributed to that organism's survival, learning, or to its ancestor's reproductive success.*

If an organism were to have no past at all – if it were to have no ancestors, and no history of past survival or learning – then it would have no task functions. Swampman, as specified, is such a creature.

On Shea's view, Swampman, at least at the moment of his creation, would lack all representational content. Despite having the same (non-intentionally characterised) causal dispositions as his normal human counterpart, Swampman – because he has no appropriate history – would have no internal representations.

Shea claims that as Swampman begins to interact with his environment, and as certain instances of his behaviour become rewarded in learning or contribute to his persistence, then a metric of 'success' or 'failure' comes to be attached to associated types of behaviour. In Shea's terminology, Swampman begins to *acquire task functions*. Once Swampman has a history of learning and persistence in the bank, his internal states become candidates to be representations (provided he exploits them in the right way to achieve behavioural success).

Notice that Swampman's internal states that are counterparts to human internal representations do not 'come online' as representations all at once. They come online gradually, and piecewise as behaviour driven by those vehicles builds up a history of reward or contributes to Swampman's persistence. It is possible that some of the representations possessed by Swampman's human counterpart *never* come online for Swampman because Swampman never has occasion to use them and/or because they do not build up an adequate history of contributing to his (post-creation) reward or persistence.

This isn't an entirely happy outcome. Let's look at a concrete case for why one might worry.

Imagine that you are a bilingual speaker of English and French. This is usually understood to require that you have many subpersonal representations related to each language (representing that language's words, phonemes, grammatical structures, etc.). Now suppose that a physical twin of you is created by a lightning bolt hitting a swamp in France. Suppose that your Swamp twin continues to live in France and never has any occasion to speak English. Despite your Swamp twin having the ability to speak fluent English, on Shea's view your Swamp twin has *no subpersonal representations associated with that ability*. This is because none of its internal states associated with its ability to speak English have contributed to its (post-creation) survival or reward. Your Swamp twin is an English speaker, but it has no associated subpersonal representations. That seems a very odd outcome; it runs counter to how linguistic abilities are normally understood. Cases like this might make you wonder whether Shea's approach is an entirely satisfactory way to handle Swampman cases.

Descriptive and directive representation

Mark Sprevak

Descriptive representations are supposed to match the way things are in the world. They are 'correct' or 'incorrect' when they do so. These types of representation are also called:

- belief-like states
- indicative states
- representations with a mind-to-world direction of fit

Directive representations are also associated with a worldly condition, but this is a worldly condition that they are supposed to bring about. They are 'satisfied' or 'unsatisfied' when they do so. These types of representation are also called:

- desire-like states
- imperative states
- representations with a world-to-mind direction of fit

In this chapter, Shea tries to pinpoint the difference between a subpersonal representation being descriptive or directive.

(NB. His account allows that certain representations can be *both* descriptive and directive. These are Millikan's *pushmi–pullyu* representations).

1 Mode + content vs content

In the wider literature on mental representation, it is common to distinguish between a representation's *mode* (aka its *attitude*) and its *content*. (For examples of this, see assigned reading by Tim Crane).

According to such a view, a representation's *content*, roughly speaking, is the state of the world that the representation is about. In contrast, the *mode* signifies the role a representation of that state of the world should play in the agent's reasoning.

For example, the two sentences:

1. The door is shut.
2. Shut the door!

have the identical content. They both represent the worldly state of the door being closed. However, they have different modes. Sentence (1) describes the door being shut; sentence (2) is a directive to shut the door (which might currently be open).

Shea departs from this standard terminology. For him, the term 'content' refers to the 'full' representational import of a state, including its mode. In other words, for him, 'content' refers to *both* the content – under the more traditional terminology – and the mode.

Shea would distinguish between (1) and (2) by saying that they have different *contents*.

Nothing important hinges on this. It is just a matter of terminology to decide what we mean by 'content'.

2 Distinction between descriptive and directive representation

It is important to bear in mind that, on Shea's view, *every* representation – irrespective of whether it is descriptive or directive – must play a causal role in generating behaviour. A subpersonal representation, *R*, must cause the organism to achieve its task function (i.e. to produce a specific form of 'successful' behaviour). More precisely stated, on Shea's view, *representation R must figure in the best causal explanation of that behaviour, B*.

For example, for the frog, the internal state of the frog that is correlated with *fly at* (x, y, z) must play a role in *causing* the frog's tongue to extend and catch that fly (the relevant behaviour that is stabilized and robust). If the frog's internal state does not cause any stabilized, robust behaviour, then it does not count as a representation on Shea's framework.

2.1 Directive representation

With this setup in mind, Shea analyses a representation with *directive* content as having an *additional* causal role.

Not only does internal representation (R) play a role in bringing about some robust, stabilized behaviour (B) for the organism, but *R also* plays a role in bringing about the worldly state (C) that it represents. More precisely stated, on Shea's view, *the fact that R brings about C* should figure in the causal explanation of how *R* produces behaviour *B*.

Think of it this way.

For every representation, *R* has to figure in the best causal explanation of *B*. More precisely stated:

(1) The fact that *R* stands in some exploitable relation – correlational information or structural correspondence – to some environmental state *C* figures in the best causal explanation of *B*.

However, for *directive* representations, two extra conditions are met:

(2) *R* brings about *C*
(3) The fact that *R* brings about *C* figures in the best causal explanation of *B*

2.2 Descriptive representation

For *descriptive* representations, the situation is slightly more complicated.

The reason it is more complicated is that, as Shea points out, it is still *possible* (although certainly not required!) that condition (2) above be true of a representation that has descriptive content. In other words, it is possible that a representation with descriptive content also brings about *C*.

To make this possibility plausible, consider motor commands. According to Shea, motor commands are *pushmi–pullyu* representations. They have both descriptive and directive content. A motor command, *R*, is both a *directive* to raise the arm (C), and also a *description* of the arm being raised (C) used by other parts of the cognitive system. Conceived of in this way, *R* has two distinct causal roles, which eventuate in two distinct sets of associated robust, stabilized behaviour:

(i) Behaviour type B_1 – causing the arm to raise
(ii) Behaviour type B_2 – changing stance of legs – a compensatory adjustment made by other aspects of the motor system caused by *R* describing that the arm is raised

R brings about B_1 (it causes the arm to raise). Hence, *R* brings about condition *C* (arms are raised in the external world). But just because *R* brings about *C*, that does not mean that *R* cannot also cause other behaviour, behaviour that does not depend on *R* having caused *C*. *R* also brings about

B_2 (it causes a change in stance of legs). The causal mechanism for R doing this does not depend on R already having caused C. R causes B_2 without 'going via' causing C to occur (in fact, timing delays in the motor system mean it would be impossible to wait until C occurs). In its role causing B_2, R functions like a descriptive representation. However, that does not change the fact that the very same state R also – *qua* its other role – happens to bring about C. R is therefore an example of an internal state with descriptive content that happens to also bring about C.

So, we cannot define descriptive content by saying that R does not bring about C.

What then should we say?

Shea argues that the crucial condition is an explanatory condition concerning whether bringing about C is, or is not, part of the causal pathway from R to the relevant behaviour. This was being hinted at in the way we handled the motor control case above.

For a *descriptive representation*, Shea says that condition (1) should be satisfied and also:

(2) If R does happen to bring about C, that fact would *not* figure in the best causal explanation of B

Returning to the motor control example, the fact that R brings about C (it causes the arms to raise) *does* figures in the explanation of B_1, but it does *not* figure in the causal explanation of B_2. When B causes B_1 the causal pathway does go through R causing C. However, when R causes B_2 the causal pathway does *not* go through R causing C (it goes nowhere near C at all). This is why, relative to R's role in causing B_1, representation R functions as a directive representation, and why, relative to R's role in causing B_2, representation R functions as a descriptive representation.

There is lots more in this chapter. Some minor tweaks are made to apply these conditions to structural representations. Shea compares his account to rival accounts of directive/descriptive distinction. He briefly discusses other types of representational content, which are neither descriptive nor directive, such as suppositional content.