Philosophy Compass 8/5 (2013): 496–509, 10.1111/phc3.12033

Naturalising Representational Content

Nicholas Shea*
King's College London

Abstract

This paper sets out a view about the explanatory role of representational content and advocates one approach to naturalising content – to giving a naturalistic account of what makes an entity a representation and in virtue of what it has the content it does. It argues for pluralism about the metaphysics of content and suggests that a good strategy is to ask the content question with respect to a variety of predictively successful information processing models in experimental psychology and cognitive neuroscience; and hence that data from psychology and cognitive neuroscience should play a greater role in theorising about the nature of content. Finally, the contours of the view are illustrated by drawing out and defending a surprising consequence: that individuation of vehicles of content is partly externalist.

The contents list is as follows:

1. The Project of Naturalising Representational Content

Some things in the world have semantic properties. Spoken and written sentences are paradigm cases. They are perfectly ordinary particulars in the causal order: ink marks on the page and vibrations in the air. But they also have more exotic properties: they can be true or false, or, in the case of imperatives, they can be satisfied, or go unsatisfied. That is, they are associated with a condition, and it makes an important explanatory difference whether that condition actually obtains, comes to obtain, or fails to obtain. Thoughts too have semantic properties in this way, and on all naturalistic views of the mind the semantic properties of thoughts are also associated with ordinary physical particulars: either whole people, or things going on within people, almost certainly involving processes in the brain.

In the 19th Century Franz Brentano identified the closely-related idea of intentionality and argued that it is a peculiar feature of thoughts (Brentano 1874/1995). Thoughts can be about objects and properties that are not present to the thinker, that are distant in time and space, that are hypothetical or may only be actualised far in the future, or that are entirely imaginary. If thinking is a physical process realised by or within people, how then can thoughts reach out and be about such things? Indeed, even when the object is right in front of me, how does my perception and thought manage to be about it, when the object is out there and the thought is in here (in some metaphorical sense)? A perfectly ordinary feature of everyday life – the fact that sentences and thoughts have

semantic properties – can come to seem mysterious and in need of explanation. There are many kinds of philosophical theory that would offer some explanatory purchase on the phenomenon, but a naturalistic account of representational content, if it were available, would dispel any sense that intentionality is mysterious or lies outside the familiar causal order of natural objects and properties.

Although some have attempted to derive the content of utterances directly from the behavioural dispositions of speakers and listeners (Millikan 2004; ch. 8; Skyrms 2010), this paper will assume that linguistic content depends at least in part on the thoughts of those communicating (Grice 1957; Lewis 1969). If so, we cannot start with public language sentences. Attempts to naturalise intentionality of all kinds will depend upon being able to naturalise mental content.

The label 'naturalising mental content' has become associated with relatively stringent assumptions about how a theory must relate mental properties to non-mental properties if it is to count as naturalistic. I simply follow that usage, leaving aside the important question of whether other accounts of how the mental relates to the non-mental also qualify as forms of naturalism (Hornsby 1997). So I will take it that a naturalistic account of mental content must provide illuminating explanatory connections between representational content and properties that are non-semantic, non-mental and non-normative. Furthermore, it must show that content properties supervene on the physical, or at least must be compatible with such supervenience. Reduction is one live option (especially for some varieties of representational content – see below), but the naturalising project is not limited to reductive theories. It is a familiar point that many special sciences deal in properties that are not reducible to basic physics. What makes them naturalistic is that their properties have substantive explanatory connections to properties in other sciences; and that they supervene on the physical. Non-reductive physicalism is compatible with the existence of *ceteris paribus* bridge laws or robust (but not exceptionless) generalisations connecting properties from different levels or schemes of explanation. Some kinds of representational content may be like that too.

This paper describes one robustly realist approach to naturalising representational content. The approach underwrites a view about the explanatory role of representational content, which is set out in section 2. Section 3 mentions some failings of existing naturalistic theories of content and section 4 draws a pluralist moral: that the metaphysics of content – what makes it the case that a representation has the content it does – may be different in different kinds of system, with no shared property that is plausibly individuative of contentful states in general. Pluralism motivates a research project that looks at a variety of systems, examining a series of predictively successful information-processing explanations in detail in order to understand what representational content must be like for those explanations to work as they do (or not – eliminativism remains an option in each case). Section 5 illustrates these views about the nature and explanatory role of content by drawing out and defending a mildly counterintuitive consequence: that the individuation of vehicles of content is a partly externalist affair.

In accordance with the brief to be opinionated, the paper is confined to one particular research strategy for naturalising representational content. While I think there are good reasons to pursue this approach, I don't take it to be the only option. On the basis of our present state of knowledge other, incompatible approaches cannot be ruled out. The research programme described here must ultimately be judged by its fruits.

2. The Explanatory Role of Content

This section sets out a widely-held form of realism about mental representation and sketches a view about the explanatory role of representational content that is compatible with it. It is the representational realism that has been deployed since the ''cognitive revolution'' in experimental psychology, cognitive neuroscience, and the other sciences of brain and behaviour. The central insight derives from the invention of mechanical computers, which gave us the idea that mental representations are physical particulars that are realized in the brains (and maybe bodies) of thinkers and interact causally in virtue of non-semantic properties (e.g. ''form''), in ways that are faithful to their semantic properties. Psychological processes like thinking, perceiving, reasoning and imagining then consist of causal processes taking place between representations with appropriate contents.

Information processing theories effectively offer a wiring diagram showing how inputs affect states of the system and, in conjunction with other states of the system, issue in behavioural outputs. What does it add to that wiring diagram to label various nodes with representational contents? A realist about mental representation is committed to the reality of the internal particulars described in the theory, and of their contents. Representationalists also typically subscribe to the claim that representations interact causally in virtue of non-semantic properties. If so, a complete causal description of the behaviour of the system is available that doesn't mention semantic properties at all. We can describe how proximal stimulation at input will cause the system to undergo various internal transitions that eventuate in movements at output. Intermediate entities at various stages of this process have semantic properties, but content does not alter the underlying causal story about internal transitions and bodily movements.

What adverting to content does achieve, however, is to show how the system connects with its environment: with the real-world objects and properties with which it is interacting, and with the problem space in which it is embedded. The non-semantic description of the system's internal organisation is true of the system irrespective of its external environment. Content ascriptions help explain how it interacts with that environment.

That can sound merely ascriptionist: that content is no more than a useful notation that makes the system comprehensible to the interpreter, with no further reality in the system. But representationalism admits of a more strongly realist treatment than that. When a system is embedded in a particular environment, entities within the system thereby acquire a suite of new, relational properties. The 'information' of information theory is a matter of covariation, so embedding a system in an environment gives rise to a host of new informational properties. An input register that could in principle carry information about very many different things now carries information about a much narrower class of conditions. Similarly there are facts about how entities in the system connect causally at output to distal objects and properties in the environment; also facts about what the system has been designed or trained, or has learned or evolved, to do. These are all perfectly un-mysterious natural facts about the embedded system. Any property that is constructed out of such properties is also instantiated when the system is embedded in an environment (automatically, as it were). If content is like that, then something more than mere ascriptionism is true. Content is a real property of the system, instantiated in part because of the way the system is embedded in its environment and in part because of the way it is internally configured – a property that is explanatory of the way the system interacts with its environment.

Instrumentalist uses of representational ascriptions are often available, and are clearly a naturalistic option. For some systems the intentional stance may even be the whole truth

about representational content (Dennett 1981), when the way a system operates cannot be broken down in terms of interacting vehicles of content. But the brand of externalist realism advocated here, when it is available, has the merit over instrumentalism of saying something more about how the system manages to perform its task, by appealing to facts about the internal operation of the system. It's not just by chance, or through a look-up table, that the system instantiates a useful input-output function; it does so by deploying an internal algorithm. The approach also goes beyond a non-semantic or purely internalist description of a system's internal operations, in that it labels the internal components with semantic contents. The external embedding is the task being performed by the system, the internal algorithm is how it manages to solve that task, and the contents of the internal items show how an internal organisation of that sort manages to perform the externally-described task. The properties appealed to in this explanation (mental contents) are more than mere convenient ascriptions. They are real properties instantiated by parts of the system. That underpins the force of appealing to these properties to explain the system's behaviour in its environment.

3. Existing Theories

That picture of the explanatory role of representational content sets an agenda for addressing, but does not itself answer, our question about the nature of content. Nor can we turn to cognitive science for the answer. Although cognitive science encompasses several hugely successful disciplines that rely on representational content as a central explanatory resource, they take the existence of semantic properties for granted. They offer no settled view about what makes it the case that the representations relied on have the contents they do. The content question has been largely left to philosophy, which has made substantial progress, especially during the 1980s.

Dretske (1981) brought the tools of mathematical information theory to bear on the issue. Information in that sense is simply a matter of correlation between types: where the state of one system changes the probability that some other system is in a given state or states. For example, an elevated firing rate in such-and-such neurons in an organism's primary visual cortex raises the probability that a horizontal line is present in its field of view. Indeed, the discovery of neurons that are differentially sensitive to such visual features by Hubel and Wiesel was one of the motivations for an information-theoretic take on representational content.

Informational accounts have to grapple with the fundamental fact that information in the correlational sense is ubiquitous. Any putative representational type carries correlational information about a whole host of natural properties, as well as about less natural rivals like disjunctive properties or Quinean collections of undetached parts. There is much more to Dretske's (1981) theory of content than that of course, but his account of which correlations are constitutive of content and which are not, and those offered by other information-based theories of content, have not yet been accepted as fully satisfactory. For example, it is often felt that they do not have an adequate answer to the 'disjunction problem' – the problem of ruling out strong putative correlations between internal types and disjunctive conditions, like *it's a fly or a moving black dot*. Nor is it at all plausible that the content of a representation is the thing with which it correlates most strongly. Exceptions abound. Proximal correlations are often stronger than the distal correlations that are more plausible candidates for content (e.g. an internal state that roughly correlates with predators but correlates more tightly with a particular pattern of shadow on the retina). Weaker correlations (e.g. indicating merely a small chance of a predator) are sometimes more plausible

candidates for content when stronger correlations are available (Godfrey-Smith 1991). Sophisticated informational theories may ultimately succeed in dealing with these difficulties (e.g. Usher 2001), but the ubiquity of correlational information is a substantial challenge for purely informational approaches to representational content.

Empirical work on the cognitive psychology of concepts also puts pressure on informational accounts. The conditions which typically cause the tokening of a given concept are many and diverse, including objects and properties with which it is connected associatively, taxonomically and thematically (Murphy 2004). Words and pictures, rather than the objects and properties themselves, are frequent causes of concept tokening. Very often the causes of the tokening of a concept are other thoughts and concepts. Jerry Fodor, who did so much precisify and elaborate the representational theory of mind, had the idea that some correlations asymmetrically depend on others (Fodor 1990). That is fine as far as it goes, but the source of the asymmetric dependence was never explained, leaving the suspicion that content was the basis of asymmetric dependence rather than the converse.

This work in cognitive psychology uncovering the relations between different concepts, and between concepts and other psychological states, points in the same direction as a long-standing idea in philosophy: that relations between representations are important to fixing their content. For example, content might depend on inferential connections between representations. As applied to concepts, that leads to conceptual role semantics, according to which the content of a concept is given by some set of dispositions the thinker has to move between it and other concepts (Block 1986). If all inferences are relevant, holism threatens (Fodor and Lepore 1992): if the identity of concept C is constituted partly by its inferential connection to concept D, whose identity in turn depends on its inferential connections to further concepts, then the individuation of C depends transitively on an entire network of interconnected concepts.

That motivates attempts to identify, for each concept, a privileged subset of dispositions that are constitutive of its content (Peacocke 1992). However, it has proven difficult to delineate sets of inferences that can do the job: that are necessary for possessing the concept, plausibly shared by most users of the concept, and sufficiently detailed to be individuative that is, to distinguish the concept from others. Cognitive psychology has shown that the dispositions thinkers actually use in deploying concepts seldom have the structure of definitions. Stored prototypes and exemplars are far more important, together with wider bodies of knowledge in some cases, like explicit or tacit theories of the domain (Margolis and Laurence 1999; Murphy 2004). For these reasons, conceptual or inferential role semantics has not so far had much success in naturalising content, except perhaps for the logical constants.

Relations amongst representations might be significant for another reason. They endow a system of representations with a structure, which may then be isomorphic to structures in the world. For example, the spatial relations amongst symbols on a cartographic map are roughly isomorphic to spatial relations amongst places on the ground; and that seems to be crucial to the way maps represent. Research on animal navigation has been a source of empirical support for the idea that isomorphisms are an important aspect of representational content (Gallistel 1990), bolstered by neural evidence about the role of the hippocampus in spatial navigation and the discovery of hippocampal 'place cells' in the rat that register the animal's spatial location (O'Keefe and Nadel 1978).[1] Cummins has done much to spell out a way isomorphism might be constitutive of representational content (Cummins 1989), but that there should be some functional isomorphism between a system of representations and the things they represent is such an extremely undemanding

condition that it has been hard to see how isomorphism, on its own, can account for the nature of content (Godfrey-Smith 1996; pp. 184–187; Shea forthcoming).

The final important step taken in the 1980s was the development of teleosemantic theories of content (Millikan 1984; Papineau 1987). Teleosemantics has two main parts. The first is to focus on how a representation is used downstream. According to teleosemantics the content of a representation is fixed in part by the way that it is used by some consumer system that reacts to or relies on the representation. The idea is to read off from the way the consumer system behaves facts about what the consumer must be taking the representation to mean. The second part of teleosemantics is to make sense of this talk of 'a consumer taking a representation to mean something' in terms of evolutionary functions. That idea applies differently to contents with different directions of fit. Indicative contents have a mind-to-world direction of fit: the content is correct when the world actually matches a specified condition. Imperative representations have a world-to-mind direction of fit: when there is a mismatch, their content is satisfied when the world is changed to match a specified condition.

The explanation of content in terms of evolutionary functions is most direct in the case of imperative representations. When a consumer system has been designed by natural selection to react to a range of different states (putative representations) with a range of different outputs, those outputs will have evolutionary functions – results that in the evolutionary past led systematically to survival and reproduction of the organism (and hence that explain why there are instances of that system around today). As well as very general conditions, for example that the output led to the organism obtaining something beneficial, there are usually outcomes specific to each type of behavioural output that the consumer produces. Empirical work on honeybee foraging provided a classic example. If a particular honeybee dance pattern is produced in the hive, it is when consumer bees fly off to a particular location to search for nectar that the behavioural output leads systematically to survival and reproduction of the hive (the functional output being to fly to a different location for each different pattern of dance). Flying to that location is therefore the imperative content of the dance.

Indicative content is given, not by output functions, but by conditions that were in place in order for those outputs to lead to survival and reproduction. As well as very general background conditions (e.g. being on the earth), again there are usually conditions that are specific to each output, and that enter into an explanation of how the output contributed systematically to survival and reproduction in the evolutionary past. Returning to the bee dance, the condition specific to a particular dance pattern is that there should be nectar at a particular location, the location dances of that type tended to dispose consumer bees to fly to in the evolutionary past. In this way indicative contents are success conditions for the behaviour of a consumer system in response to a representation – where success is naturalised in terms of performing an evolved function.

Teleosemantics helped cut down some of the liberality of correlational information, but it faced worries of its own about founding content on facts about a system's history. A molecule-for-molecule duplicate that was created by chance would behave in just the same way as an evolved system but, according to teleosemantics, it would have no states with representational content (Braddon-Mitchell and Jackson 1997). It also remained unclear how well the theory handles high-level representational states like beliefs and desires, although the extension to functions based on learning may help here (Dretske 1988; Millikan 1984). Teleosemantics derives considerable support from empirical work on animal signalling, like the honeybee nectar dance (von Frisch 1967) and the remarkable experiments on signalling by vervet monkeys about different kinds of predator

(Seyfarth et al. 1980). In the case of human beliefs and desires, we do not yet have a clear psychological account of how these representations are consumed in downstream processing. If the 'consumption' of beliefs and desires is a general purpose affair, for example being a matter of how they feed into theoretical and practical inference, then it is less clear how an appeal to functions will ground the attribution of conditions, specific to each belief, for the performance of its specific functions (in an evolutionary normal way).

4. Pluralism

So the recent history of attempts to naturalise representational content is a story of many ideas and no conclusive resolution. Every view faces serious problems as a full and unified theory of content. Nor is there consensus about which approach is most promising. The overview pointed at ways in which different lines of empirical evidence motivate and support different approaches to content. The empirical cases lead in different directions. In this section I want to suggest that a moral can be drawn from this rather partisan story: that there may be no one true unified account of the nature of content. The metaphysics of content may be different in different kinds of representational system.

In a paper on folk psychology Peter Godfrey-Smith argues that folk explanations of behaviour in terms of mental representations might work for a variety of different reasons in different cases (Godfrey-Smith 2004). He also observes that cognitive scientists differ radically amongst themselves as to which of the naturalistic properties discussed above – correlation, isomorphism, inferential roles, teleofunctional specificity – are most fundamental to semantic content. He draws a pluralist conclusion: that different semantic concepts are suited to different circumstances (p. 160).

I want to adopt Godfrey-Smith's insight. Put in terms of the realist approach to representation set out in section 2 above, different kinds of representational property might be suited to explaining the behaviour of different kinds of system. These different representational explanations share a commitment to explaining how a system performs some externally-specifiable task (a function from distal inputs to external outcomes) in terms of the interaction of various internal components, with content capturing the way those components are connected to external circumstances in ways that are useful for performing the externally-specified task (by correlation, isomorphism, etc.). The relevant connections or 'exploitable relations' (Godfrey-Smith 2006) may be different in different cases, as may the other external factors that make it the case that the system is a representational system at all. If that is right, pluralism about representational content follows, at least in subpersonal systems. There may be no one true unifying theory. What makes it the case that a representation has the content it does may be different in different kinds of representing system.

If that is right, then there is a simple diagnosis of why the rapid philosophical progress on naturalising representational content ground to a halt. Different theories drew support from different bodies of empirical literature, but none could cover all the cases because the very nature of the representational content needed to explain the behaviour of different kinds of system is different.

Given pluralism about content, we should look at representational explanations of behaviour in a wide variety of different domains, in order to uncover a variety of kinds of representational content. Where an information processing theory is successful at predicting behaviour and supported by evidence about internal states, that raises a *prima facie* case that the representations relied on are real, and have the contents relied on in the theory. That is not to say that philosophers have to take scientists' theories on trust.

The scientists may be wrong about what contents are being processed, or about whether their explanations need appeal to representational content at all. However, we can investigate whether a given predictively-successful explanation of behaviour really does rely on representational contents. If it does, that gives us a defeasible reason to think that those representations in that system have those contents. It is then a defeasible constraint on a theory of content that, as applied to that system, it should deliver those contents. That is a way of arriving at contents for a system that is not just based on intuition. Over-reliance on trading intuitions about subpersonal systems like frog retinal ganglion cells was an unfortunate feature of some earlier philosophical debates about content.

To be more precise, the strategy I am advocating is to examine a variety of representational explanations, and for each to identify:-

(a) An explanandum concerning how the system operates or behaves in relation to its environment.
(b) A putative explanation of (a) that relies in part on attributing representational properties to the system (e.g. keeping track of p, aiming at q, etc.).
(c) An account of how the explanation in (b) succeeds (remaining open to there being no such account).
(d) If there is a positive answer to (c), a characterisation of the kind of properties the representational properties of the system would have to be for the explanation in (b) to succeed in explaining (a) in accordance with the account (c).

For example, there is a well-confirmed account of the information processing responsible for subjects' behaviour when they make a series of rapid choices between options which are only rewarded probabilistically. Subjects learn by reinforcement and there is now strong behavioural and neural evidence that they deploy what is called a temporal-difference learning algorithm (Schultz et al. 1997). This is a field where neural data has played an important role in deciding between theories. Rival theories postulate different internal states intermediate between input and behaviour. Mathematical models tell us the quantitative values these intermediates should take, which vary from trial to trial. These quantitative intermediates can be very different according to different information-processing theories, even when their behavioural outputs are equivalent. The technique of model-based analysis looks for corresponding trial-by-trial variation in data about neural firing rates, most often obtained indirectly from the fMRI BOLD signal (Corrado et al. 2009). This new way of making use of fMRI is considerably more powerful than the standard subtractive method in uncovering the information processing responsible for various simple forms of human behaviour, as well as delivering detailed knowledge of how those representations are realised in the brain. Temporal-difference learning accounts of reward-guided decision making have had considerable predictive success and have been very influential in cognitive neuroscience and neuroeconomics. The representational contents they rely on therefore offer a good target for philosophical theories of content (Shea 2012).

To follow this pluralist strategy philosophers will have to look to a wider range of data from the sciences of brain and behaviour than they have to date. We saw above how important empirical data has been in motivating different theories of content; data for example on the correlational information carried by different neural areas, cognitive maps in the medial temporal lobe, ways that people deploy concepts and functional animal signalling. If we cast the net even more widely we may uncover further insights about the types of features that are relevant to content.

The examples discussed above concern subpersonal level representations (although some cases do have implications for the personal level: Shea 2013). Personal level contents are available to consciousness and enter into epistemic or reason-giving relations for the whole person. Subpersonal level contents are not and do not, although they can also underpin intelligent behaviour. Pluralism about the nature of content as between personal level representations like beliefs and desires and the several varieties of subpersonal level contents is especially plausible. Nor should we expect content to be the same kind of thing in conceptual representations, which have content-bearing constituents, as it is in non-conceptual representations, which have correctness conditions but lack semantically-significant constituent structure.

Why, then, should we expect an understanding of content in subpersonal systems to have any relevance for personal level content? The strategy I advocate is to tackle the project of naturalising personal level content from below. We start with the simpler cases to learn more about how content-based explanations work and to catalogue a variety of content-relevant things that are going on in different systems. The hope is that these insights will allow us to creep up on the problem of personal level content gradually, gaining an understanding of increasingly complex systems which have more moving parts and increasing interactions with personal level phenomena. The complexities of personal level phenomena like believing and desiring may look more tractable when we have a better understanding of simpler representational systems with which they are likely to share many features, especially if the personal-subpersonal distinction is not an unbridgeable ontological or explanatory divide.

5. Externalist Syntax

To illustrate the way internal and external factors conspire to give rise to content according to this form of externalist realism, I will draw out and defend a mildly revisionary consequence of the framework: a form of externalism about syntax. Whether a system has the kind of realist externalist contents discussed above depends in part on its environmental embedding. That is in tension with the idea that there is a syntactic description of the internal causal operation of a representational system that is completely independent of features of the environment. Amongst other things, a syntactic description enumerates the vehicle types. A consequence of the view expressed above is that whether a causal description of the operation of a system counts as a *syntactic* description depends in part on the system's environment. Often a reasonably complex system is suited to implementing more than one algorithm or computational process, without altering the internal workings of the system. Which of those counts as a syntactic description will depend on which task environment the system is embedded in. That runs contrary to a widespread assumption that, even if semantic externalism is tenable, syntax is a wholly internalist affair.

For example, Fodor claims that 'computational [states] are individuated by their local properties' (and are syntactic items: Fodor 1994, p. 14). Internalist syntax was also presupposed in the debate over whether the explanatory properties relied upon by psychology are or must be externalist (Davies 1991; Egan 1991; Segal 1991). A few authors since have questioned that assumption (discussed below), but other than those who adopt the vehicle externalism of the extended mind hypothesis (Clark and Chalmers 1998), internalism about syntax remains the orthodoxy. The radical vehicle externalism of the extended mind hypothesis is not the issue we are considering here, because our vehicles of content are entities that are parts of the system in question, in a way that extended mind hypothesis rejects. Our question is rather whether the fact that an internal state is a

syntactic item at all could in principle depend upon which environment the system is located in.

There may be strong constraints on which kinds of systems could be implementations of particular syntactically-described operations (cf. Sprevak 2010), for example that could implement the first-order predicate calculus. But we are working with a conception of representational processing that encompasses any system that makes transformations between vehicles in a way that is faithful to their semantic properties, where 'faithfulness' can include inductive and abductive inferences, and inferences that reflect merely local statistical regularities. Very many physical systems are capable of implementing some such transitions. Conversely, a reasonably complex physical system is capable of implementing content-faithful transitions in more than one way (example below). Rather than saying vehicles of content are absolutely ubiquitous, the framework above makes it a substantive constraint on being a syntactic item that an internal representation stands in the right relations to the system's environment to bear (externalist) content.

An example motivates the idea that a system could be divided up into vehicles of content in more than one way, with the appropriate way depending in part on the context in which the system is embedded (Shagrir 2001 provides another example). Consider a system that operates as shown in Fig. 1. It takes areas or blobs as input and transforms them into areas at output. For example, the inputs and outputs could be illuminated areas on a touch screen. We assume the inputs and outputs are sufficiently fine-grained that we can treat them as continuous; or we could imagine the system implemented by physically manipulating volumes of fluid, which would be practically continuous. Suppose too that the way the system is connected to its task environment is appropriate to make it the case that its internal intermediates have content.

The operations the system performs are as follows. The input screen is divided into two different sections, Input 1 and Input 2. The system counts the number of shapes in Input 2 and divides the area of each shape in Input 1 into that number of parts. It does the converse to each shape in Input 2. That produces the two intermediate stages illustrated in Fig. 1, which both have a number of shapes equal to the number in Input 1 multiplied by

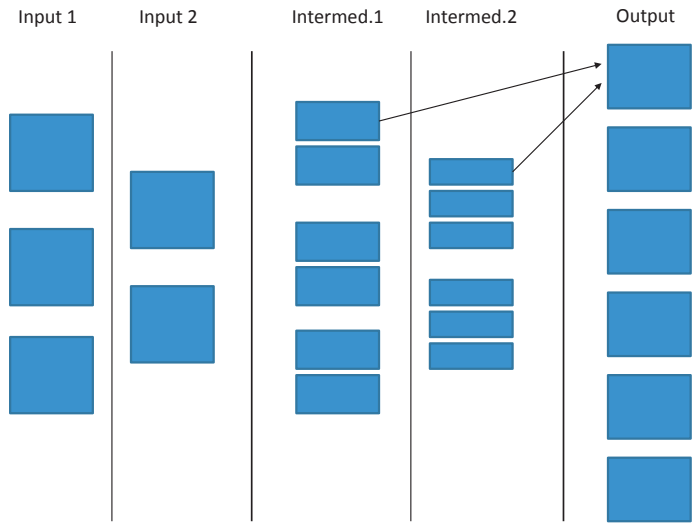

Fig. 1. A hypothetical system with at least two possible syntactic descriptions.

the number in Input 2. Finally, the system merges each shape in Intermed. 1 with a shape in Intermed. 2 to form an output shape with the combined area of the two.

The system is contrived so that there are two equally-good ways of describing its operation. Most obviously, it multiplies the number of shapes in Input 1 by the number of shapes in Input 2 and represents the product by the number of shapes at Output. Less obviously, it adds the total surface area at Input 1 to the total surface area at Input 2 and represents the result by the total surface area at Output. Which calculation is the system performing?

According to the framework offered here, the answer depends on how the system is embedded in a problem space. Perhaps it is being used to calculate the amount of ink one would need in order to print the two patterns of shapes presented as input, with the colour density delivered at output providing an at-a-glance representation of the total ink required. Or it could be used as a simple touch-based multiplier to teach multiplication to young children. In the first case the system is an adder; in the second a multiplier. Both inputs and outputs differ in content depending on the case: the outputs represent continuous volume in the first case and discrete integers in the second. The representational content of the intermediate states also varies. In the first case, Intermed. 1 represents the total area of Input 1, and Intermed. 2 the total area of Input 2. In the second case, Intermed. 1 represents the product of the numbers of shapes in Input 1 and Input 2, as does Intermed. 2.

In this example, it is not just that different contents arise in the two different embeddings. The very way states of the system are divided up into vehicles of content differs between the two cases. The divisions between different blobs, which is crucial to the system's implementing multiplication, is irrelevant to the way the system adds areas. And the fine-grained differences in area which constitute different vehicles in the adder all count as instances of the same vehicle in the multiplier if they fall within the same blob. That is, the appropriate syntactic description of the system's operation depends on the problem space in which it is embedded.

Although it is still standardly assumed that vehicle individuation and syntax are intrinsic to a system, the kind of moderate externalism about vehicles advocated here follows an established line of thought. In particular, Shagrir (2001) draws on the kinds of consideration set out above to argue that content externalism implies what he calls computational externalism: that which syntactic structures are implemented by a system depends in part on factors external to the system. (Horowitz 2007) follows Shagrir in arguing for what Horowitz calls 'wide computational properties'. (Crane 1990) argues that internal states cannot be syntactic unless they also have semantics, against some computation-based accounts of the individuation of syntax, but does not go on to draw externalist conclusions. (Bontley 1998) argues that the appropriate syntactic description of a system may depend on its design or teleofunction, so he subscribes to an historically-based externalism about syntactic types.

How does this square with the commonly held view that the individuation of vehicles of content should depend only on local properties, not on facts external to the system (Fodor 1994, pp. 14–15)? Can representationalism survive the loosening of that view? The additional commitment that distinguishes representationalism from instrumentalism is that vehicles of content should be real material entities, that interact in virtue of their non-semantic properties. What explanatory advantage does that commitment secure? As we've seen, it allows us to explain how a system manages to implement a given input-output function, by showing that it implements an algorithm that breaks that function down into a series of steps. Without realism about the vehicles described in the

algorithm, the algorithm is no more than an instrumentally-useful way of describing the input-output function being performed.

When the system does indeed have an internal organisation that corresponds to an algorithm for performing an input-output function, an additional predictive and explanatory benefit follows. We get to predict a certain kind of stability and characteristic pattern of change. Representations are likely to be gained and lost piecemeal. So we can explain the system's representational stability over time in terms of the persistence of internal vehicle types which bear those contents; and we can explain the piecemeal gain and loss of representations (like acquiring a new belief) in terms of the internal change implementing a corresponding change to vehicle types. Furthermore, an error at one step in the algorithm is likely to affect all and only downstream processing that relies on that step.

Vehicle individuation need not be entirely internalist to secure these advantages. What representationalism requires is that vehicle types correspond to real internally-specifiable types. That is consistent with the view that *which* internally-specifiable types are vehicle types depends in part on factors external to the system; in particular, in our case, on which division into internal vehicle types corresponds to an algorithm for performing the input-output function in which it is embedded. Representationalism only requires that there should be intrinsic properties of the system that sort tokens into vehicles. Where the contentful explanation of the system's behaviour appeals to the same representation, the syntactic account should have vehicles with the same internally-specifiable properties. Accordingly, the kind of moderate externalism about vehicle individuation advocated here is consistent with securing the explanatory benefits that realism about representation has over instrumentalism.

6. Conclusion

Progress towards naturalising representational content may be achieved by aiming to account for the representational contents presupposed by well-confirmed scientific theories that appeal to information-processing to explain behaviour. In pluralist spirit, we should not assume that the nature of content will be the same in every case. Although we constrain our theories of content through the lens of the explanatorily useful, that is consistent with representations being real internal entities interacting in virtue non-semantic properties, and contents being real, partly-relational properties of those vehicles. The resulting moderate externalism about vehicle individuation still secures the explanatory benefits of realism about representational content.

Acknowledgement

The author would like to thank Tim Bayne, Ruth Millikan, Dan Ryder, Mark Sprevak and Ulrich Stegmann for discussion and comments on earlier drafts; the audience at the Oxford philosophy of mind work-in-progress group for helpful discussion; and Ron Mallon and an anonymous referee for comments on a previous draft. This work was supported by: the Wellcome Trust (grant number 086041), the Oxford Martin School and the John Fell OUP Research Fund.

Short Biography

Nicholas Shea is an interdisciplinary philosopher of mind, and of psychology, cognitive science and cognitive neuroscience. Following a PhD at King's College London, he worked

in the Faculty of Philosophy at the University of Oxford, before returning to King's. As well as philosophical work on mental representation, consciousness, inheritance systems and the metaphysics of mind, he has published in scientific journals in collaboration with psychologists, cognitive neuroscientists and biologists. His current focus is on the nature of representation in relatively low-level psychological systems.

Notes

* Correspondence: Department of Philosophy, King's College London, Strand, London WC2R 2LS, UK. Email: nicholas.shea@kcl.ac.uk.

[1] Hippocampal place cells are evidence of correlation rather than isomorphism, but combined with other evidence on spatial navigation abilities and the role of the hippocampus and wider medial temporal lobe, it has been suggested that these cells play a role in forming a wider 'cognitive map' that is indeed isomorphic to spatial features of the environment.

Works Cited

Block, N. 'Advertisement for a Semantics for Psychology.' *Midwest Studies in Philosophy, Vol. 10: Studies in the Philosophy of Mind*. Eds. P. A. French, T. Uehling, H. Wettstein. Minneapolis: University of Minnesota Press, 1986. 615–78.

Bontley, T. 'Individualism and the Nature of Syntactic States.' *British Journal for the Philosophy of Science* 49 (1998): 557–74.

Braddon-Mitchell, D. and F. Jackson. 'The Teleological Theory of Content.' *Australasian Journal of Philosophy*, 75.4 (1997), 474–89.

Brentano, F. C. *Psychology From an Empirical Standpoint*. London: Routledge. 1874/1995.

Clark, A. and D. C. Chalmers. 'The Extended Mind' *Analysis*, 58 (1998): 7–19.

Corrado, G. S., et al. 'The Trouble With Choice: Studying Decision Variables in the Brain.' *Neuroeconomics: Decision Making and the Brain*. Eds. P. W. Glimcher, C. F. Camerer, E. Fehr, R. A. Poldrack. Amsterdam: Elsevier, 2009. 463–80.

Crane, T. 'The Language of Thought: No Syntax Without Semantics.' *Mind & Language* 5 (1990): 187–212.

Cummins, R. *Meaning and Mental Representation*. Cambridge, MA: MIT Press. 1989.

Davies, M. 'Individualism and Perceptual Content.' *Mind*, 100.400 (1991), 461–84.

Dennett, D. C. 'True Believers: The Intentional Strategy and Why It Works.' *Scientific Explanation*. Ed. A. F. Heath. Oxford: OUP, 1981. 53–75.

Dretske, F. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press. 1981.

Dretske, F. *Explaining Behaviour: Reasons in a World of Causes*. Cambridge, MA: MIT Press. 1988.

Egan, F. 'Must Psychology be Individualistic.' *The Philosophical Review* 100.2 (1991): 179–203.

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

—— *The Elm and the Expert*. Cambridge, MA: MIT Press: Bradford. 1994.

—— and E. Lepore. *Holism: A Shopper's Guide*. Oxford: Wiley-Blackwell. 1992.

von Frisch, K. *The Dance Language and Orientation of Bees*. Oxford/Cambridge, MA: OUP/Harvard University Press, 1967.

Gallistel, C. R. *The Organization of Learning*. London/Cambridge, MA: MIT Press. 1990.

Godfrey-Smith, P. 'Signal, Decision, Action.' *Journal of Philosophy* 88 (1991): 709–22.

—— *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press. 1996.

—— 'On Folk Psychology and Mental Representation.' *Representation in Mind: New Approaches to Mental Representation*. Eds. H. Clapin, P. Staines, P. Slezak. Amsterdam: Elsevier, 2004. 147–62.

—— 'Mental Representation, Naturalism and Teleosemantics.' *New Essays on Teleosemantics*. Eds. D. Papineau, G. Macdonald. Oxford: OUP, 2006. 42–68.

Grice, P. 'Meaning.' *Philosophical Review* 66 (1957): 377–88.

Hornsby, J. *Simple Mindedness: A Defence of Naïve Naturalism in the Philosophy of Mind*. Cambridge, MA: Harvard University Press. 1997.

Horowitz, A. 'Computation, External Factors, and Cognitive Explanations.' *Philosophical Psychology* 20.1 (2007): 65–80.

Lewis, D. *Convention*. Cambridge, MA: Harvard University Press. 1969.

Margolis, E., & Laurence, S., eds. *Concepts: Core Readings*. Cambridge, MA: MIT Press, 1999

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

—— *Varieties of Meaning*. London/Cambridge, MA: MIT Press. 2004.
Murphy, G. L. *The Big Book of Concepts*. London/Cambridge, MA: MIT Press. 2004.
O'Keefe, J. and L. Nadel. *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press. 1978.
Papineau, D. *Reality and Representation*. Oxford: Blackwell. 1987.
Peacocke, C. *A Study of Concepts*. Cambridge, MA: MIT Press. 1992.
Schultz, W., P. Dayan, and P. R. Montague. 'A Neural Substrate of Prediction and Reward'. *Science*, 275.5306 (1997): 1593.
Segal, G. 'Defence of a Reasonable Individualism.' *Mind*, 100.400 (1991): 485–94.
Seyfarth, R. M., D. L. Cheney, and P. Marler. 'Vervet Monkey Alarm Calls: Semantic Communication in a Free-Ranging Primate'. *Animal Behaviour*, 28.4 (1980): 1070–94.
Shagrir, O. 'Content, Computation and Externalism.' *Mind*, 110.438 (2001): 369–400.
Shea, N. 'Reward Prediction Error Signals are Meta-Representational.' *Noûs* (2012). doi:10.1111/j.1468-0068.2012.00863.x
—— 'Neural Mechanisms of Decision Making and the Personal Level.' *Oxford Handbook of Philosophy and Psychiatry*. Eds. K. Fulford, M. Davies, G. Graham, J. Sadler, G. Stanghellini, T. Thornton. Oxford: Oxford University Press, 2013.
—— 'Millikan's Isomorphism Requirement.' *Millikan and Critics*. Eds. J. Kingsbury, D. Ryder, K. Williford. Oxford: Wiley-Blackwell, forthcoming.
Skyrms, B. *Signals: Evolution, Learning, & Information*. Oxford/New York: OUP, 2010.
Sprevak, M. 'Computation, Individuation, and the Received View on Representation.' *Studies in History and Philosophy of Science*, 41 (2010): 260–70.
Usher, M. 'A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation.' *Mind & Language*, 16.3 (2001): 311–34.