Representational Development Need Not Be Explicable-By-Content

Nicholas Shea

Abstract

Fodor's radical concept nativism flowed from his view that hypothesis testing is the only route to concept acquisition. Many have successfully objected to the overly-narrow restriction to learning by hypothesis testing. Existing representations can be connected to a new representational vehicle so as to constitute a sustaining mechanism for a new representation, without the new representation thereby being constituted by or structured out of the old. This paper argues that there is also a deeper objection. Connectionism shows that a more fundamental assumption underpinning the debate can also be rejected: the assumption that the development of a new representation must be explained in content-involving terms if innateness is to be avoided.

Fodor has argued that connectionism offers no new resources to explain concept acquisition: unless it is merely an uninteresting claim about neural implementation, connectionism's defining commitment to distributed representations reduces to the claim that some representations are structured out of others (which is the old, problematic research programme). Examination of examples of representational development in connectionist networks shows, however, that some such models explain the development of new representational capacities in non-representational terms. They illustrate the possibility of representational development that is not explicable-by-content. Connectionist representations can be distributed in an important sense, which is incompatible with the assumption of explanation-by-content: they can be distributed over non-representational resources that account for their development. Rejecting the assumption of explanation-by-content thereby opens up a more radical way of rejecting Fodor's argument for radical concept nativism.

Contents

For *Fundamental Issues of Artificial Intelligence*, ed. Vincent C. Müller (Synthese Library).

(1) Introduction

How can new representations be acquired? When that question is asked about new concepts, Fodor famously argued that hypothesis testing is the only option (Fodor 1975, 1981). That led him to embrace radical concept nativism.[1] Many objectors have pointed out that hypothesis testing is not the only candidate learning mechanism (Carey 2009, Cowie 1999, Laurence & Margolis 2002), showing how existing representations can be involved in the acquisition of new ones without the new representations thereby being structured out of the old (Margolis 1998, Rupert 2001).[2] This paper argues that the point runs deeper. The development of a new representation need not be explicable in content-involving terms at all. It may proceed by putting together non-representational resources in such a way as to constitute an entirely new representation.

Most answers to Fodor's challenge rely on existing representational resources in explaining the development of a new representation type. For example, a new natural kind term can be acquired by recognising salient properties of an object, connecting them with a new internal symbol, and combining that with an essentialist principle (Margolis 1998). Such accounts are important in their own right, but do not challenge the tacit assumption that, if innateness is to be avoided, representational development must consist in a series of stages or transitions that are explicable in terms of the semantic content of the representations involved. Connectionist systems offer an excellent illustration of the more radical claim. Fittingly, Fodor's robust challenge to the usefulness of connectionist modelling brings out the central role that an assumption of explanation-by-content is playing in his arguments.

One of Fodor's challenges to connectionism is characteristically pithy: connectionists' defining characteristic is a commitment to distributed representations — over *what* are they distributed? At best, Fodor argues, connectionists can be saying nothing more than that connectionist representations are distributed over further representations, those found at the level of their processing units (section 2). But if so, connectionism is nothing new. It is just a restatement of the old programme in which new representations are complex constructs out of innate primitives. For present purposes we can concede to Fodor that that programme has been unsuccessful, especially as an account of lexical concepts.

Fodor's challenge forces us to be very precise about what connectionism should say about the development of new representations. This paper examines some examples to show that even static feedforward PDP models provide a genuinely novel way of seeing the development of new representational resources. They show us how there can be a non-representational but informational explanation of the development of entirely new

---

1 Fodor has since retreated somewhat from that position: Fodor (1998, 2008), including becoming more sympathetic to alternatives to hypothesis testing as an account of concept acquisition (Fodor 2008, pp. 162-168).

2 Fodor now accepts that such processes do not depend on hypothesis testing (Fodor 2008), but still argues that they form part of a creature's innate conceptual endowment (2008, pp. 163-164).

primitive representations (section 3). Connectionism's interesting distribution claim is that representations can be distributed over the entities that account for their development (section 4). That is perhaps the most important theoretical insight offered by PDP modelling.

At first sight, Fodor's challenge to connectionism appears to be independent of his argument for radical concept nativism. On examination, they both turn out to depend upon the same tacit assumption: that if the development of a new representation is to fall within the ambit of psychology, it must occur as a rational transition or inference between semantic items, explicable in virtue of their contents (it must be 'explicable-by-content'). Connectionist models of representational development show that there can be a psychological explanation of the transition to a new representation that is not an explanation-by-content. That point generalises into a tactic for answering Fodor's puzzle of radical concept nativism in other cases (section 5). However, the nativists had a good point too, because it would be a mistake to think, as many connectionists have, that the acquisition of new representational resources is to be explained in terms of their contents. The grain of truth in Fodor's radical concept nativism is that, for very many of the representations that are responsible for intelligent behaviour, their development is not explicable-by-content. His mistake was to conclude that they must therefore be innate.

Fodor's innateness argument concerns concepts. Concepts are one species of mental representation. They are constituents of complete thoughts. Complete thoughts have associated conditions of satisfaction or truth/correctness. Concepts, taken alone, do not. Fodor hypothesises a language of thought, such that all mental representations with conditions of truth/correctness or satisfaction are formed out of constituent concepts. Representationalism is more permissive allowing, for example, that there are representations which have satisfaction conditions that we would express using a sentence (eg, *there is a snake on the ground, climb a tree*), but where the representation itself contains no constituent structure (nothing in the representation corresponds separately to *the ground, snakes* or *climbing*). I will use the term 'non-conceptual representation' for psychological states with correctness or satisfaction conditions but no constituent structure. Non-conceptual representations are probably needed to understand many classes of PDP model, as well as many psychological phenomena. Although Fodor talks about concepts, the considerations he canvasses in support of nativism are equally applicable to non-conceptual representations. So I will move freely between talking about concepts and representations in general.

(2) Fodor's Argument Against Connectionism

Fodor has several objections to connectionism. The most well-known is that connectionist models cannot explain the systematicity and productivity of thought (Fodor and McLauglin

1990). Also prominent is the claim that connectionists cannot avoid an unacceptably holist theory of the content of distributed representations (Fodor and Lepore 1992). My focus is a third objection: that connectionism is nothing new. According to Fodor, all it has to offer is the standard idea that some mental representations are structured out of others, coupled with an outdated associationism about mental processing (Fodor and Pylyshyn 1988). Supposed theoretical insights, like the idea of distributed representation, and of learning by modulation of connection strengths, are simply new pieces of terminology for old ideas, terminology that obscures the failings inherent in treating lexical concepts as structured, but does nothing to address them.

The argument can be formulated as a dilemma.[3] Connectionists claim that mental representations are distributed. What type of objects are they distributed over? Characteristically, Fodor offers a dichotomy:

Version One: Mental representations are distributed over neurons.[4]
Version Two: Some mental representations are distributed over others.[5]

According to Version One, connectionism is just a claim about how mental representations are realised. Any psychological theory must be realised somehow in physical brains. No one thinks that all mental representations correspond to individual neurons or "grandmother cells". To be of theoretical interest, connectionists need to make some claims that connect with the explanatory level of psychology. So Fodor argues.

PDP modellers themselves are unlikely to accept the Version One characterisation. Of course when the models are applied to the real world, representations will be realized in multiple neurons. But the units over which representations are distributed in the models are not neurons. PDP modellers often explicitly eschew a commitment to a 1-1 correspondence between processing units and neurons. Version One therefore does not capture the force of PDP's distribution claim.

Fodor offers Version Two as the only alternative. But this is just a familiar story about structured representations. Connectionists' distributed representations are merely some kind of complex constructs out of the representations that are their constituent units. For example, Fodor and colleagues interpret Churchland's "state space semantics" (Churchland 1998, 2012) as treating individual hidden layer units as representing complex microfeatures, with a distributed pattern of activation having its content as some kind of complex weighted conjunction of these microfeatures (Fodor and Pylyshyn 1988, § 2.1.4; Fodor and Lepore 1999, p. 391). Distributed representations are structured out of the

[3] This objection to both the versions of connectionism offered here has been raised by Fodor in many places, e.g. in Fodor and Pylyshyn (1988) and Fodor and McLaughlin (1990). The formulation explicitly in terms of a dilemma is found in Fodor (2004), a draft paper posted on the New York University website.

[4] Fodor and Pylyshyn (1990), § 5.6.

[5] Fodor and Pylyshyn (1990), § 2.1.4.

representations over which they are distributed, and the content of a distributed representation is fixed by the contents of the constituent units.

Fodor rejects Version Two on the basis that constructing concepts out of pre-existing representations is a failed research programme. He argues that there are no plausible definitions for most lexical concepts, and that neither prototypes nor exemplars compose in the way that is required by a compositional semantics. Connectionists can, and often do, object at this stage. Fodor's objections to prototype and exemplar theories may be surmountable. Or the connectionist's way of constructing distributed representations out of the contents of individual nodes may be different in important respects, so as to overcome extant objections. Furthermore, "constructivist" neural networks side-step the worry about constructing distributed representations out of existing microfeatures since they allow for the recruitment of new hidden units that previously played no role in the network (Mareschal & Schultz 1996, Quartz & Sejnowski 1997).

These lines of reply to Fodor are familiar. They may be the best way to characterise some classes of connectionist models. But there is another answer available too. Fodor's dilemma presupposes that there are only two candidates for the entities over which mental representations are distributed: neurons or further representations. To have any bite connectionists do indeed have to tell us what it is that distributed representations are distributed over. But Fodor has offered connectionists a false dichotomy. To see that there is another possibility we must first get on the table a positive account of the development of new representations in connectionist systems.

(3) Developing New Connectionist Representations

Even static, programmer-designed neural networks can develop novel representations. This section gives an account of how. In particular, it shows how familiar training algorithms can transform a system without representations into one that has representational capacities.

The basic idea is that there are connectionist learning algorithms that transform information into representation. Before a connectionist system has been trained, the units of its hidden layers, and perhaps its input layer too, can be merely information-carriers. Their tokening will correlate with various features of the items coded as input. When the instantiation of some property F by an object changes the probability of the instantiation of another property G by an object, we can say that F carries *correlational information* about G. Correlational information is ubiquitous. Representation is something more substantial. The fact that single units and distributed patterns of activation carry correlational information (about all sorts of affairs) does not imply that they have representational content. Typically, it is only after training that distributed patterns of activation have the right properties to have genuinely representational content (truth conditions, satisfaction

conditions, etc.). Of course, there is no agreement as to exactly what more is needed, but all sides agree that bare correlational information is not sufficient for representation.

Connectionists need not think of individual units as being representational at all. Indeed, Chalmers (1992) takes that to be characteristic of connectionist models: the items over which computational processes are defined are more fine-grained than the lowest level at which representational contents are properly attributable to states of the system. (Some networks *are* designed to have individual units as representational, e.g. the semantic networks of Quillian 1967.) Connectionists have tended to accept that individual units represent something (e.g. complex microfeatures), when they need not. That takes connectionism towards the Version Two interpretation and its attendant problems. In fact, in many networks there is no reason to think of individual units as representational at all.

An example is the colour classification network of Laakso and Cottrell (2000). One way of coding the inputs there proceeded as follows. For a given colour patch, reflectance readings from a spectrophotometer were taken at 12 places on the electromagnetic spectrum (between wavelengths of 400nm and 700nm, at 25nm intervals). The readings were normalised to the range 0-255 and converted into binary format (eg, 11010011), giving a list of 12 binary numbers for each colour sample. This list of binary numbers was converted into a 96-dimensional vector of 0s and 1s to act as input vector (96 = 12 binary numbers of 8 digits). In that coding it is very hard to see an individual one of the 96 input units as representing anything at all. The particular 0 or 1 it carries makes sense only as part of a binary representation of magnitude that is distributed across 8 units. So even at the input layer there are cases where the individual units are not representational and only distributed patterns of activation are.

The case is even clearer when we come to hidden layers. There are good reasons, in many classes of model, not to treat single units of a hidden layer as representational. It is a mistake to concede that individual hidden layer units represent some kind of complex microfeatures. Shea (2007b) describes a class of connectionist systems in which individual hidden layer units are not representational. The networks do feature distributed representations. However the representations are distributed over network units, not over further representations. In other cases, the representations may be dynamic attractors in activation space (Clark 2001, p. 135, e.g. McLeod et al. 2000), making the representational level even further removed from the individual units in a single layer. Importantly, an explanation of how new representations develop (in those cases, how clusters or dynamic attractors develop in hidden layer state space) *is* given at the level of individual units.

So individual units in a connectionist network may not be representations: individual hidden layer units are unlikely to have representational content before training, and in many cases individual input layer units do not have representational content at all. But notice that each individual unit will carry correlational information, both before and after training (indeed, units will carry information about very many properties of the samples that have been coded into inputs). In many connectionist networks, training encourages

the network to form representations.[6] Some simple examples bring out the point. Competitive networks use unsupervised learning to find clusters in the inputs on which they are trained (Rolls and Treves 1998, ch. 4). Unsupervised learning in auto-associative networks can also serve to identify the central tendencies or prototypes found in a range in input data, even where the prototype itself was never encountered in training (Plunkett and Sinha 1992). Both start with bare correlational information and end up with vehicles (clusters, prototypes) that are plausibly representations.

The point about the development of new representations can be made most starkly in networks in which there is no representation at all at the level of individual units before training, like the examples above. But that is not essential. The absence of initial representations just serves to make it obvious that the way that new representational capacities develop is not explicable-by-content. Representational development in these cases is a matter of using statistical learning to build mere information-bearers into representations. In other cases, pre-existing resources that are representations play a role in this process. What is crucial is that their role is merely causal. The way a new representational type develops, at a hidden layer say, depends on the correlational information carried by input units and hidden units, but there is no rational or content-based explanation of the transition from initial resources to new representations. Existing resources like the input units are relied on for the correlational information they carry, on which the connectionist training algorithm can act; but the story of the building of the new representational capacities is causal-correlational, not representational.

Application to an Example

To discuss a widely-known example, Sejnowski and Rosenberg's (1987) NETtalk network was trained using supervised learning to map English text to phonetic representations of its pronunciation. Where networks undergo supervised learning, clusters may form in hidden layer state space, leading to new distributed representations at the hidden layer. In NETtalk, before training there were no relevant partitions or clusters in hidden layer state space (although distributed patterns of activation would necessarily have carried some correlational information from the outset). The result of training the network to produce correct representations of phonemes at the output layer was that the network learnt to categorise inputs into vowels and consonants at the hidden layer on the way.

According to two representative theories of content, asymmetric dependence theory and infotel semantics, this process leads to the creation of new representations out of non-representational resources. Learning a new representation of Cs is a matter of acquiring a new mental item R with the right properties firstly, to count as a mental symbol, and secondly to have the content C. According to Fodor's asymmetric dependence theory of

[6] There are many examples in which learning in connectionist systems creates attractors or clusters in state space (Churchland & Sejnowski 1992, Rupert 1998, 2001, Tiffany 1999). If there are reasons to see those attractors as being representations, then this is a process of turning information into representation.

content, having a representation R with content C is a matter of having a mental symbol whose tokening covaries with the presence of Cs, and of asymmetric dependence: to the extent that the tokening of R also covaries with any other property C*, it would not so-covary if R did not also covary with C (Fodor 1990). Call the mechanism which puts R in the right relation of causal covariation and counterfactual dependence with C a *sustaining mechanism* (Cowie 1999, p. 101; Laurence & Margolis 2002). Fodor's theory of content has faced many objections, and Fodor doesn't seem particularly keen on it himself,[7] but taking it at face value, learning a new representation of C is just a matter of going through a psychological acquisition process which results in a sustaining mechanism that connects a new symbol type R with Cs (with the appropriate causal profile).

Applying the asymmetric dependence theory to NETtalk, it is reasonably clear that there is no representation in the hidden layer at the outset, when connection weights are set randomly or arbitrarily. From the outset, both input and hidden layer units will carry a variety of correlational information, but there is no basis for thinking that there are any relations of asymmetric dependence amongst these correlations. After training, activation of the vowel partition of hidden layer state space correlates with presentation of a vowel to the network. It also correlates with other properties of the stimulus, say with the stimulus being a letter with a certain disjunctively-specified shape S. But that correlation is plausibly asymmetrically dependent on the correlation with vowels — were it not for the correlation with vowels, which is a useful intermediate to the classification made at the output layer, the network would not have arrived at a correlation with shape S.

For contrast, we can also assess the representational contents in NETtalk using infotel semantics (Shea 2007a), a modification of teleosemantics (Millikan 1984; Papineau 1987). Infotel semantics looks at the way a representation is used, as well as the way it is produced, in fixing its content. Out of all the correlational information carried by a putative representation, it focuses on the correlation that accounts for the system's having been trained (or evolved) to behave as it does (as argued by Dretske 1998; Ryder 2004 deploys a related idea). Applied to PDP models, this will deliver as content a condition specific to each representation-type, such that keeping track of that condition is what enables the network to produce correct outputs (where correctness is the standard against which the learning algorithm was trained).

Applied to NETtalk, infotel semantics implies that the output layer represents phonemes: in the course of training, the modeller took the units to represent phonemes, using that as the standard against which to generate an error signal. At the hidden layer, before training we have only correlational information. After training we have a partition of activation space into two groups of distributed patterns. Each correlates with a relevant feature of the input (vowel vs. consonant), and that distinction is consumed in downstream

[7] 'I assume that intentional content reduces (in some way or other, but, please, don't ask me how) to information; this is, I suppose, the most deniable thesis of my bundle.' (Fodor 1994, p. 4). 'If you want an externalist metaphysics of the content of innate concepts that's not just bona fide but true, I'm afraid there isn't one "yet".' Fodor (2001), p. 137.

processing as a means to further phonetic categorisation. So according to infotel semantics there are representations in the hidden layer after training, but there is only correlational information before.

Notice that in this case, although there are representations at the input layer throughout (of strings of text), their role in fixing the content of the new representations (vowel vs. consonant) formed at the hidden layer is merely causal. It is not as if *vowel*, say, has been defined as some complex property of strings of text. Instead, input encodings of words into text strings serve as the causal basis for a sustaining mechanism that connects clusters at the hidden layer with properties of words. There is no explanation-by-content of the transition from a non-representational hidden layer, before training, to representations of vowels and consonants at the hidden layer, after training.

(4) Connectionism's Interesting Distribution Claim

Armed with this account of the way connectionist networks can develop novel representations out of non-representational resources, we can return to Fodor's dilemma: *what* are connectionist representations distributed over?

In cases like the hidden layer representations in NETtalk, distributed representations are the lowest level of grain at which representational contents are properly attributable to the system. This fits Chalmers' (1992) observation that computational processes go on at a more fine-grained level (individual units) than the lowest level at which representations are found (distributed patterns of activation). Similarly, the story about how new representations develop is located at the more fine-grained level of single units. It is a recognisably psychological story, a form of statistical learning based on the way activation of units correlates with external features, and on correlations in activation between units.

We can clearly distinguish between the two levels of grain: one at which representations are found, another which figures in an account of the development of new representations.[8] That is, we can distinguish between ways of carving the network up into individuals for two different purposes:-

Obj1. Vehicles of representational content — individuals that figure in a representational explanation of the synchronic online operation of the trained system.

Obj2. Developmental units — individuals that figure in an explanation of the development of new representations.

Fodor's argument against connectionism assumes that Obj1=Obj2. But connectionists can make a much more interesting claim: that Obj1 are distributed over Obj2. The objects over which representations are distributed are not further representations (Version Two

[8] Tiffany (1999) and Shea (2007b) make parallel claims about the vehicles of content.

connectionism). Nor are they something merely implementational like neurons (Version One connectionism), since Obj2 are individuals which *do* figure in a psychological explanation (of the development of new representations). The interesting connectionist claim is that representations can be distributed over the resources that lead to their development. That is a clear sense in which connectionists' commitment to distributed representations *is* something new. It breaks away from an assumption that is deeply entrenched in classical computational models — that development of new representations must take place over existing representational resources. In this way, connectionist modelling has furnished cognitive science with a genuine insight, opening up a previously unexplored portion of logical space.

There would be good reason to assume Obj1 = Obj2 if we were committed to the idea that the development of a new representation must be explained in contentful terms: as an inference or rational transition that makes sense in the light of the semantic content of the objects involved in that transition. That is to reject the possibility that individual units may have a causal role in the development of a new distributed representation in virtue of the correlational information they carry, not constituting its content directly, but instead forming a sustaining mechanism which gives rise to its content. That is, Fodor's dichotomy implicitly assumes that the development of a new representation must be explicable-by-content:

> Assumption of explanation-by-content
> Whenever it occurs by a psychological process, the development of a new representation must consist of a transition from existing representational resources to the new representation, explicable in terms of the contents of the respective representations.

When new representations develop in PDP models in the way analysed in the previous section, it is clear that Obj1 are distributed over Obj2. The identification of Obj1 with Obj2 is a substantive assumption that has been implicitly constraining theorising in cognitive science. It is motivated by the assumption of explanation-by-content. That assumption also underpins Fodor's strong innateness claims, as we shall see in the next section.

(5) Avoiding Fodor's Argument For Radical Concept Nativism

But what about Fodor's argument for radical concept nativism? We have seen how even static connectionist models can account for the development of entirely novel representations. They are not innate: PDP models offer an account of their development,

and does so in recognisably psychological terms.[9] How, then, is Fodor's argument avoided?

At first pass, Fodor's argument that connectionism offers nothing new seems quite separate from his argument for radical concept nativism. In this section we will see that Fodor's argument for radical concept nativism is in fact underpinned by the same assumption that lay behind his identification of Obj1 with Obj2 in the last section. Connectionism's insight is to show why that assumption can be rejected. In this section we spell out how doing so side-steps Fodor's nativism puzzle.

Fodor argues that concepts are either constructed from primitives or they are innate. His view is that most lexical concepts — concepts at the level of grain of individual words — are not constructed out of primitives. So they are innate, which is to say that they are not acquired via a learning process. Why does Fodor think that learning can only consist in constructing new representations out of existing ones? Not all ways of acquiring a new representation count as learning. Neither a bump on the head nor clever neurosurgery are learning processes, so if new representations could be acquired in either of those ways they would not be learnt. By contrast, setting a parameter for a grammatical principle, detecting a correlation, and constructing a new prototype based on experience are all clear cases of learning. Fodor argues that they all involve testing a hypothesis about what is the case: that the ambient grammar is head-first, that A correlates with B, that birds typically have feathers. To test a hypothesis against experience, the learner has first to be able to represent the hypothesis. So hypothesis testing cannot be a way of acquiring genuinely new representational resources, ones whose expressive power extends beyond contents that can be constructing out of pre-existing representations.

The standard response is that not all learning mechanisms are forms of hypothesis testing. That answer is correct, but it is incomplete, because it doesn't tell us what learning processes look like that are not hypothesis testing.[10] Fodor's move equating learning with hypothesis testing is not just an observation about what learning happens to consist in. It goes deeper. The claim is that learning can only consist in rational transitions between representations (Fodor 1975, p. 36). If that were right, then a person would indeed need to be able to formulate a claim before they could learn that it was true, which would exclude a learning-based account of the acquisition of entirely novel representations. That presents a puzzle, since it is implausible that my concepts of a carburettor (CARBURETTOR) or of my friend John (JOHN) are innate.[11]

---

9 The concept of innateness is notoriously problematic (Mameli 2008). Fodor's central concern is whether concepts are learnt (Fodor 1975, 1991, 1998, 2008; Cowie 1999; Samuels 2002), so here I will take it that innate representations are not learnt or acquired by a psychological process and that they admit of a poverty of the stimulus argument (Shea 2012a, 2012b).

10 Margolis (1998), Rupert (2001), Laurence & Margolis (2002) and Carey (2009) give detailed accounts of forms of concept learning that are not a matter of hypothesis testing.

11 Fodor has softened slightly in more recent work. First he allowed that concepts themselves may not be innate – what is innate is, for each concept, a domain-specific disposition, specific to each such concept, to acquire that concept (Fodor 1998). But this still leaves Fodor postulating an innate domain-specific ability to develop DOORKNOB as a result of interaction with doorknobs. He has since added that the innate endowment might determine the geometry of neural attractor landscapes that realise concepts (Fodor 2008, p. 164). The worry remains that far too much is being taken to be innate. For simplicity, this paper

Several authors have suggested a strategy for answering Fodor's innateness puzzle (Macnamara 1986, Margolis 1998, Rupert 2001, Laurence & Margolis 2002). Assume that what makes a representation have the content it does is wholly or partly determined by its causal relations with things in the world. Such sustaining mechanisms for a representation R may depend, causally, on other representations R* without the content of R being determined by the content of the R* — R's content is fixed more directly by its causal relations with things in the world. R can then be atomic, neither structured nor constructed out of the R*. The anti-nativist tactic is to give a psychological story in which existing representations R* come to form the sustaining mechanism for a new representation R, where the process of forming the new representation type R is described merely causally, not as a content-driven process like inference.

The key to this strategy is that not all learning consists in rational transitions between representations. Fodor's commitment to explanation-by-content closes off that option (driving him toward innateness). And we can see why Fodor would think learning is restricted in that way. The central insight of cognitive science is the viability of content-based explanation — the explanation of behaviour in terms of rational transitions between mental representations. The reality of these mental processes is vindicated by causal transitions between representation tokens in virtue of their form, but explanatory purchase is achieved by describing such representations in terms of their content. Rational transitions between contentful representations are the very core of the representational theory of mind (and its offshoot, the computational theory of mind / language of thought). So it is natural that Fodor should think that all psychological processes must consist in transformations between mental representations that are explicable in terms of the content of those representations.

If all learning processes were like that, then Fodor would be right to claim that any way of acquiring new representations that did not relate them to existing representations would necessarily lie outside the explanatory ambit of psychology. We have seen a first response to Fodor in accounts where the development of new representations depends upon existing representations without the new representation being structured or constructed out of the old (Margolis 1998, Laurence & Margolis 2002, Rupert 2001). But we can go further and reject the deeper underpinnings of Fodor's argument if we can reject the assumption of explanation-by-content entirely.[12] We must show how there can be instances of learning that are susceptible to a recognisably psychological explanation, but which do not fit within the standard mould where the outcome (a new representation) can be explained as a rational transition from existing resources. The transition to a genuinely

considers only Fodor's earlier innateness claim.

12 Fodor has more recently accepted that these accounts of concept learning do not involve hypothesis testing (Fodor 2008, pp. 163-167), and even that there is a 'jump' from the existing representations that are involved in creating the prototype: 'we jump, by some or other "automatic" process, from our stereotypes to our concepts' (2008, p. 164). However, he does not draw the moral that there are psychological acquisition processes that are not explicable-by-content; indeed, he argues that the way this process works is due to innate constraints (2008, p. 164).

novel contentful item cannot itself be susceptible to explanation in terms of content.

Our account of the development of new representations in the PDP models in section 3 above is an existence proof that there can be such cases. It escapes Fodor's argument for radical representational nativism by rejecting his implicit commitment to explanation-by-content. That commitment can now be seen to lie behind both his radical concept nativism and his rejection of connectionism. But once PDP modelling has opened up this portion of logical space, it becomes clear that other cases of representational development should be understood in the same way. Shea (2011) has argued that Carey's influential account of children's development of the concept of natural number (Carey 2009) also involves a step that is not explicable-by-content.[13] Below I offer an example that goes beyond connectionism to illustrate that this could be a more general phenomenon.

Face Recognition

Morton & Johnson's (1991) theory of the development of face recognition furnishes a further useful example of how acquisition could fashion representations out of purely non-representational resources. Tested 30 minutes after being born, infants show a tendency preferentially to look at moving stimuli that have a configuration something like this:

This tendency seems to be innate, in the sense that no learning is involved in the infant coming to have the looking bias. A poverty-of-the-stimulus argument can be made about it. The infant's disposition preferentially to track this category of inputs (perhaps driven by a subcortical visuomotor pathway) implicitly carries the information that such stimuli are worth attending to and learning about. If we ask where *that* information came from, we have to appeal to the infant's evolutionary history, not its individual experience. We can suppose that the bias is adaptive — it works well in the kinds of environments infants are likely to find themselves in. The adaptive match between behavioural bias and usual environment is due to evolution, not individual learning.

The infant's unlearnt behavioural bias is then sufficient to give a second system the input it needs to learn to reidentify individual faces. Through being given the right kind of input, this learning system has the chance to extract the statistical properties that distinguish one face from another and the statistical invariants that signify the same face again. Once trained up, the second system also implicitly encodes information: a rich store of information about which features indicate the same face (John, say). Unlike the information in the initial visual tracking tendency, this latter match between system and

[13] Carey also observes that the child makes a 'leap' when drawing a parallel between the operation of adding one object in the object file system and the process of counting on to the next item in the (initially uninterpreted) sequence of counting words. Shea (2011) argued that this is the step at which Fodor's argument is circumvented, and that this step is not explicable as a rational transition from the content of pre-existing representational resources.

environment is not due to evolution, but to individual learning (from the experience of seeing John).

What contents are represented at these two stages of development? The answer depends upon the correct theory of content, about which there is no consensus, so I will again deploy asymmetric dependence theory and infotel semantics. At birth infants have the capacity to detect moving blobs and certain configurations of blobs. Some internal state driving their looking behaviour covaries roughly with the presence of faces, but there do not seem to be asymmetric dependencies between the various kinds of information carried or, if there are, it is the capacity to detect faces that looks to be asymmetrically dependent on the capacity to detect configurations of blobs, rather than the other way round. So, according to Fodor's theory of content, infants do not represent faces at the outset.[14] As a result of learning, the infant comes to be able to reidentify a particular individual, John say, by his face: the infant categorises together a variety of different views of John, and can engage in John-relevant behaviour as a result. So the result is some internal vehicle which correlates with John, and may well have the right asymmetric dependence properties to count as a representation of John. Thus, according to asymmetric dependence, the infant initially has no representation of faces at all, but then comes to have the ability to represent John by his face.

Infotel semantics also delivers the result that the capacity to represent John is not innate. Since the visuomotor tracking bias present at birth seems to have the function of enabling learning about faces, it plausibly carries the content *that's a face, look at it*, even though it is only able to identify faces very roughly at that stage. So infotel semantics suggests that this basic capacity to represent faces is innate. The capacity to represent the particular individual John is not innate. Although, even at birth, there are features of the visual signal that correlate with the presence of John, these are not deployed by consumer systems in a John-relevant way. Only once learning has taken place, so that the infant can reidentify John and thereby engage in John-relevant behaviour, will infotel semantics deliver any representations of John. Thus, according to both theories of content, the capacity to represent John is not present initially, but only arises after the second system has done its job.

We have offered a psychological account of the development on the ability to represent John, but the transition to having a representation of John is not explicable-by content, whether asymmetric dependence or infotel semantics is the right theory of content. The capacity to represent lines and blobs figures only causally in the development of the sustaining mechanism for the infant's later representation of John. Having played its developmental role in selecting appropriate input, the initial visual tracking tendency plays no causal role in the synchronic operation of the sustaining mechanism (Johnson et al. 1991).

---

[14] Since it is not clear how the relevant counterfactuals are to be assessed, it is hard to reach definitive conclusions about how the theory will apply to specific cases.

The developmental transition is not explained-by-content. Instead, it makes use of resources characterised in terms of correlational information, which do not have the characteristics needed to count as representational, and builds them into sustaining mechanisms that do count as representational (according to both asymmetric dependence and infotel semantics). Even before learning, the visual signal carries information, in the correlational sense, about particular faces. There is something about the visual signal which correlates with looking at John, say — that is why statistical learning about individual faces works. On no view does this correlational information, present in some complex form in the visual signal, count as representational at the initial stage. But these information-bearers play a causal role in the development of the mature ability to reidentify John. The story of that developmental transition is an account of information-bearers being built up into a sustaining mechanism for a new representation. It is a psychological story, but it is not explanation-by-content. It is a psychological account of the creation of content.

Why is this developmental transition an instance of learning, rather than mere triggering or maturation? Because it is a psychological process that involves extracting information from the environment. The infant comes to represent a particular individual, John, by interacting with John. If that were just triggering, it would be a mere accident that causal intercourse with John was needed to trigger maturation of the infant's concept of John.[15] By contrast, according to Morton and Johnson's theory there is a very obvious reason why the ability to recognise John depends upon seeing John: because the learning process works by picking up statistical properties in visual signals that come from John — properties that go with its being the same face again.

One way of confirming that the mature representation of John is not innate is by deploying a poverty of the stimulus argument. A poverty of the stimulus argument is available about the neonate's ability selectively to track things which tend to be faces. So it is plausibly innate. But there is no poverty of the stimulus argument available about the infant's later ability to recognise John. The infant's John-recognition device implicitly encodes a wealth of information about which properties distinguish John's face from other faces and which properties are invariant over different views of John's face. The infant does not rely on its evolutionary history for that information (and could not), but extracts that information from its experience of interacting with John. So if Morton and Johnson are right, the infant's capacity to represent individual faces is not innate. Whether or not they are right, their theory provides a detailed example of how genuinely novel representations could be learnt.

In the last section we saw that PDP modelling of representational development opens up a new portion of logical space for cognitive science to explore: that representations are distributed over the resources that account for their development, breaking the link

[15] Fodor says that interaction with doorknobs is needed to trigger the DOORKNOB concept because being a doorknob is a response-dependent property Fodor (1998). Whether or not that response works for DOORKNOB, it is implausible that being John (a particular person) is a response-dependent property.

between psychological explanation and explanation-by-content. In this section we saw that the same tactic gives us a more general answer to Fodor's puzzle about representational innateness.

(6) Conclusion

Fodor assumes that all psychological processes, including concept acquisition, are intentional, i.e. explicable-by-content. All processes that are not susceptible to intentional explanation are bundled together under the label 'innate'. That puts acquiring the concept JOHN or CARBURETTOR by rich interactions with John or with actual carburettors on a par with acquiring such concepts via an accidental bump on the head. But Fodor has given us a false dichotomy. It is a familiar point that Fodor's model of concept acquisition as hypothesis testing is too restrictive. The further point is that representational acquisition need not be explicable-by-content at all, but may still be recognisably in the domain of psychological explanation.

This paper has shown that to be more than just a theoretical possibility. Connectionist models offer concrete examples of that process. In order to see connectionist models as accounting for the development of new representations we have to reject the assumption that development of new representations can be explained-by-content, an assumption that lies at the heart of Fodor's critique of connectionism, and of his radical concept nativism. This paper argues that we should reject that assumption and embrace the idea that some connectionist models show how new primitive representations can develop in response to the environment, without relying on pre-existing representational resources. Connectionism's answer to the question, 'Over *what* are your representations distributed?' opens up a new portion of logical space for cognitive science. Connectionist representations can be distributed over the objects that figure in an account of their development. In that way, connectionist modelling has provided a deep philosophical insight and an important contribution to theoretical progress in cognitive science: new, non-innate representations can develop in ways that are not explicable-by-content.

Acknowledgements

The author would like to thank the follow for generous comments: David Braddon-Mitchell, Steve Butterfill, Nick Chater, Martin Davies, Jerry Fodor, Paul Griffiths, Peter Godfrey-Smith, Richard Holton, Matteo Mameli, David Papineau, Gualtiero Piccinini, Kim Plunkett, Paul Smolensky, Mark Sprevak, Scott Sturgeon; and audiences in Melbourne, Oxford and Sydney.

References

Carey, S. (2009), *The Origin of Concepts* (Oxford / New York, O.U.P.)

Chalmers, D. (1992), 'Subsymbolic computation and the Chinese room', in J. Dinsmore (ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap*, (Hillsdale, NJ, Lawrence Erlbaum).

Churchland, P. M. (1998), 'Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered', *Journal of Philosophy* 95(1), 5-32.

Churchland, P. M. (2012), *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals* (MIT Press).

Churchland, P. S. and T. J. Sejnowski (1992), *The Computational Brain*. Cambridge, Mass: MIT Press.

Clark, A. (2001). *Mindware*. Oxford: O.U.P.

Cowie, Fiona (1999), *What's Within?* (Oxford: OUP).

Dretske, Fred (1988), *Explaining Behaviour: reasons in a world of causes* (Cambridge, MA: MIT Press).

Fodor, J. A. (1975), *The Language of Thought* (London / Cambridge MA, MIT Press).

Fodor, J. A. (1981), 'The present status of the innateness controversy', in *Representations: philosophical essays on the foundations of cognitive science* (London / Cambridge MA, MIT Press).

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

Fodor, J. A. (1998), *Concepts: Where Cognitive Science Went Wrong* (New York, OUP).

Fodor, J. A. (2001), 'Doing without what's within: Fiona Cowie's critique of nativism', *Mind,* 110, 99-148.

Fodor, J. A. (2004), 'Distributed representations; enough already',
http://www.nyu.edu/gsas/dept/philo/courses/representation/papers/fodordistributed.pdf accessed 2 June 2014.

Fodor, J. A. (2007), 'The revenge of the given', in McLaughlin & Cohen (eds.), *Contemporary Debates in Philosophy of Mind* (Oxford, Blackwell).

Fodor, J. A. (2008), *LOT 2: The language of thought revisited* (Oxford / New York, O.U.P.).

Fodor, J. A. & E. Lepore (1992), *Holism: A shopper's guide*. Oxford: Blackwell.

Fodor, J. A. & E. Lepore (1999), 'All at sea in semantic space: Churchland on meaning similarity', *Journal of Philosophy* 96(8), 381-403.

Fodor, J. A. and B. McLaughlin (1990). 'Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work.' *Cognition* 35: 183-204.

Fodor, J. A. and Z. W. Pylyshyn (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis.' *Cognition* 28: 3-71.

Johnson, M.H., et al. (1991). 'Newborns' preferential tracking of face-like stimuli and its subsequent decline'. *Cognition*, 40, 1-19.

Laakso, A. & G. Cottrell (2000), 'Content and cluster analysis: assessing representational

similarity in neural systems', *Philosophical Psychology* 13(1), 47-76.

Laurence, S. and E. Margolis (1999), 'Concepts and Cognitive Science' in *Concepts: Core Readings* (Cambridge, MA, MIT Press).

Laurence, S. and E. Margolis (2002), 'Radical Concept Nativism', *Cognition* 86, pp. 25-55.

Macnamara, J. (1986), *Border dispute: the place of logic in psychology* (Oxford / New York, O.U.P.)

Mameli, Matteo (2008), 'On innateness: the clutter hypothesis and the cluster hypothesis', *Journal of Philosophy*, 55, 719-36.

Margolis, Eric (1998), 'How to acquire a concept', *Mind & Language*, 13 (3), 347-69.

Mareschal, D., & Schultz, T. R. (1996). 'Generative connectionist networks and constructivist cognitive development'. *Cognitive Development*, 11, 571-603.

McLeod, P., Shallice, T. and Plaut, D. C. (2000). 'Attractor dynamics in word recognition: converging evidence from errors by normal subjects, dyslexic patients and a connectionist model', *Cognition*, 74, 91-113.

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

Morton, J., and Johnson, M.H. (1991). 'CONSPEC and CONLEARN: a two process theory of infant face recognition'. *Psychological Review*, 98, 164-181.

Papineau, D. (1987). *Reality and Representation*. Oxford: Blackwell.

Plunkett, K. and Sinha, C. (1992), 'Connectionism and Developmental Theory', *British Journal of Developmental Psychology*, 10(3), pp. 209-254.

Quartz, S. R., & Sejnowski, T. J. (1997). 'The neural basis of cognitive development: A constructivist manifesto'. *Behavioral & Brain Sciences*, 20, 537–596.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5), 410-430.

Rolls, E. and A. Treves (1998). *Neural Networks and Brain Function*. (Oxford, OUP).

Ryder, D. (2004), 'SINBAD neurosemantics: A theory of mental representation', *Mind & Language*, 19 (2), 211-40.

Samuels, Richard (2002), 'Nativism in Cognitive Science', *Mind & Language*, 17, 233-65.

Shea, N. (2007a). 'Consumers Need Information: supplementing teleosemantics with an input condition', *Philosophy and Phenomenological Research*, 75(2), 404-435.

Shea, N. (2007b), 'Content and Its Vehicles in Connectionist Systems', *Mind & Language* 22(3), pp. 246-269.

Shea, N. (2011) 'New concepts can be learned', review essay on Susan Carey, *The Origin of Concepts*. *Biology & Philosophy*, 26, pp. 129-139.

Shea, N. (2012a), 'Genetic representation explains the cluster of innateness-related properties', *Mind & Language*, 27 (4), 466–93.

Shea, N. (2012b), 'New thinking, innateness and inherited representation', *Philosophical Transactions of the Royal Society B*, 367, 2234-44.